SCIENCE AND TECHNOLOGY TEXT MINING: HYPERSONIC AND SUPERSONIC FLOW

BY

DR. RONALD NEIL KOSTOFF
OFFICE OF NAVAL RESEARCH
800 N. QUINCY ST.
ARLINGTON, VA 22217
PHONE: 703-696-4198/ FAX: 703-696-4274
INTERNET: KOSTOFR@ONR.NAVY.MIL

MR. HENRY JOSEPH EBERHART
NAVAL AIR WARFARE CENTER CHINA LAKE (RET'D)
CHINA LAKE, CA

MR. DARRELL RAY TOOTHMAN
RSIS, INC.
MCLEAN, VA

(*THE VIEWS IN THIS PAPER ARE SOLELY THOSE OF THE AUTHORS AND DO NOT REPRESENT THE VIEWS OF THE DEPARTMENT OF THE NAVY, ANY OF ITS COMPONENTS, OR OF RSIS, INC*)

ABSTRACT

Database Tomography (DT) is a textual database analysis system consisting of two major components: 1) algorithms for extracting multi-word phrase frequencies and phrase proximities (physical closeness of the multi-word technical phrases) from any type of large textual database, to augment 2) interpretative capabilities of the expert human analyst. DT was used to derive technical intelligence from a hypersonic/ supersonic flow (HSF) database derived from the Science Citation Index and the Engineering Compendex. Phrase frequency analysis by the technical domain expert provided the pervasive technical themes of the HSF database, and the phrase proximity analysis provided the relationships among the pervasive technical themes. Bibliometric analysis of the HSF literature supplemented the DT results with author/ journal/ institution publication and citation data. Comparisons of HSF results with past analyses of similarly structured near-earth space and Chemistry databases are made. One important finding is that many of the normalized bibliometric distribution functions are extremely consistent across these diverse technical domains.

KEYWORDS: HYPERSONIC; SUPERSONIC; HIGH MACH NUMBER; HIGH SPEED FLOW; SHOCK WAVE; BOUNDARY LAYER; DATABASE TOMOGRAPHY; TEXT MINING; BIBLIOMETRICS; CITATION ANALYSIS; LOTKA'S LAW; BRADFORD'S LAW

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
17 NOV 2003		**-**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
SCIENCE AND TECHNOLOGY TEXT MINING HYPERSONIC AND SUPERSONIC FLOW	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
Ronald Kostoff; Henry Eberhart; Darrell Toothman;	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
Office of Naval Research,800 North Quincy Street,Arlington,VA,22217,	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Database Tomography (DT) is a textual database analysis system consisting of two major components: 1) algorithms for extracting multi-word phrase frequencies and phrase proximities (physical closeness of the multi-word technical phrases) from any type of large textual database, to augment 2) interpretative capabilities of the expert human analyst. DT was used to derive technical intelligence from a hypersonic/ supersonic flow (HSF) database derived from the Science Citation Index and the Engineering Compendex. Phrase frequency analysis by the technical domain expert provided the pervasive technical themes of the HSF database, and the phrase proximity analysis provided the relationships among the pervasive technical themes. Bibliometric analysis of the HSF literature supplemented the DT results with author/ journal/ institution publication and citation data. Comparisons of HSF results with past analyses of similarly structured near-earth space and Chemistry databases are made. One important finding is that many of the normalized bibliometric distribution functions are extremely consistent across these diverse technical domains.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**		**46**	

1. INTRODUCTION

Science and technology are assuming an increasingly important role in the conduct and structure of domestic and foreign business and government. In the highly competitive civilian and military worlds, there has been a concommittent increase in the need for scientific and technical intelligence to insure that one's perceived adversaries do not gain an overwhelming advantage in the use of science and technology. While there is no substitute for direct human intelligence gathering, there have become available many techniques which can support and complement direct human intelligence gathering. In particular, techniques which identify, select, gather, cull, and interpret large amounts of technological information semi-autonomously can augment and expand greatly the capabilities of human beings for performing technical intelligence.

One such technique is DT [Kostoff, 1991, 1992, 1993a, 1994b, 1995], a system for analyzing large amounts of textual computerized material. It includes algorithms for extracting multi-word phrase frequencies and phrase proximities from the textual databases, coupled with the topical expert human analyst to interpret the results and convert large volumes of disorganized data to ordered information. Phrase frequency analysis (occurrence frequency of multi-word technical phrases), performed by the topical domain expert, provides the pervasive technical themes of a database. Phrase proximity (physical closeness of the multi-word technical phrases) analysis provides the relationships among pervasive technical themes, as well as among technical themes and authors/ journals/ institutions/ countries, etc. The present paper describes use of the DT process, supplemented by literature bibliometric analyses, to derive technical intelligence from the published literature of HSF science and technology.

In particular, HSF, as defined by the authors for this study, consists of hypersonic (flows with characteristic Mach Numbers greater than about five) and supersonic (flows with Mach Numbers ranging from about one to five) flow over aerodynamic bodies. Hypersonic and supersonic free jet expansions into low pressure regions, a major hypersonic/ supersonic theme in the literature, was included insofar as it supported the understanding of flows over aerodynamic bodies. Use of supersonic beams for chemistry experiments (where the gas dynamics of the beam was not the primary interest but rather served to enable the chemistry resulting from beam interactions) was not included.

To execute the study reported in this paper, a database of relevant HSF articles was generated using the unique iterative search approach of Simulated Nucleation [Kostoff, 1997a]. Then, the database was analyzed to produce the following characteristics and key features of the HSF field: recent prolific HSF authors; journals which contain numerous HSF papers; institutions which produce numerous HSF papers; keywords most frequently specified by the HSF authors; authors whose works are cited most frequently; particular papers and journals cited most frequently; pervasive themes of HSF; and relationships among the pervasive themes and sub-themes.

What is the importance of applying DT and bibliometrics to a topical field such as HSF? The roadmap, or guide, of this field produced by DT and bibliometrics provides the demographics and a macroscopic view of the total field in the global context of allied fields. This allows specific starting points to be chosen rationally for more detailed investigations into a specific topic of

interest. DT and bibliometrics do not obviate the need for detailed investigation of the literature or interactions with the main performers of a given topical area in order to make a substantial contribution to the understanding or the advancement of this topical area, but allow these detailed efforts to be executed more efficiently. DT and bibliometrics are quantity-based measures (number of papers published, frequency of technical phrases, etc.), and correlations with intrinsic quality are less direct. The direct quality components of detailed literature investigation and interaction with performers, combined with the DT and bibliometrics analysis, results in a product highly relevant to the user community.

The remainder of the report is structured as follows: Section 2 provides a background literature survey of roadmaps, and the use of co-occurrence phenomena for information retrieval and research evaluation. The reader familiar with this literature can proceed directly to section 3, which describes the database generation for the present study. Section 4 presents the study results, section 5 contains the study conclusions, and section 6 is the bibliography.

2. BACKGROUND

2.1 Overview

This section shows the unique features of the computer and co- word-based DT process relative to other roadmap techniques. It describes the two main roadmap categories (expert-based and computer-based), summarizes the different approaches to computer- based roadmaps (citation and co-occurrence techniques), presents the key features of classical co-word analysis, and shows the evolution of DT from its co-word roots to its present form.

2.2 Overview of Roadmaps

The relationships among science and technology fields, and the temporal evolution of these relationships, have been of long-term interest to many organizations. The 'roadmaps' of these relationships have been used for science and technology marketing; for science and technology management including planning, executing, reviewing, and transitioning; for enhancing communications among researchers, technologists, managers, users, and stakeholders; for identifying gaps and opportunities in science and technology programs; for technical intelligence; and for identifying obstacles to rapid and low-cost product development. The generalized roadmap relates science and technology performed at some point in time to: its science and technology heritage; other relevant science and technology being performed at the same time; and future relevant science and technology and eventual end products.

2.3 Major Roadmap Categories

There have been many roadmap studies reported in the literature [e.g., Zurcher, 1997; Groenveld, 1997; Gill, 1994; ARPA, 1994; Barker, 1995]. With minor variations, there are two fundamental approaches which are employed: expert-based and computer-based.

1) Expert-Based Approach

The expert-based approach draws on the knowledge and experience of teams of experts to subjectively identify the structural relationships within the science-technology-mission network and to subjectively specify the quantitative and qualitative attributes of the links and nodes which comprise the roadmap network.

Some of the more well-known formal expert-based prospective techniques include PERT [Dodin, 1985; Lootsma, 1988], GERT [Seyedghasemipour, 1987], and Critical Path Method [Arsham, 1993]. These approaches tend to be more applicable to optimizing project management, where the component technologies and desired end targets and schedules are fairly well understood. For nominal prospective science and technology modeling, where relationships among the network nodes can have high uncertainty, less formal network constructs are probably the most widely used [Groenveld, 1997; Barker, 1995].

2) Computer-Based Approach

In most of the techniques placed in this category, large textual databases which describe science, technology, engineering, and end products are subject to computer analyses. These databases could include published papers, reports, memoranda, letters, etc. Through the use of generic computerized methodologies including computational linguistics and citation analyses, research, technology, engineering, and product areas are identified; their relative importance is estimated and quantified; their relationships and linkages to other areas are identified and quantified. Once all these node and link attributes have been specified, a network of relationships can then be constructed.

In contrast to the expert-based approach, the computer-based approach has a more objective basis. The resultant roadmap is derived 'bottom-up' from the literature. It does not have the preconceived limitations, constraints, biases, and personal and organizational agendas of the experts. The computer based approach is in its infancy, due to the recent emergence of large textual databases with the relevant information, and recent development of computational linguistic approaches which can extract and order the database material to provide useful information. The computer-based approach is the basis for the present paper.

2.4 Types of Computer-Based Approaches for Roadmaps

There are two major types of computer-based approaches for developing the science-technology-mission structural relationships. One type tracks publication and patent citation lineages [e.g., Narin, 1989; Perko, 1997] to develop data for the relational maps. While relationships among science and technology areas are obtained, the basic citation data portrays temporal evolution. The other type, which includes DT, exploits the use of co- occurrence phenomena. In co-occurrence analysis, phenomena that occur together frequently in some domain are assumed to be related, and the strength of that relationship is assumed to be related to the co-occurrence frequency. Networks of these co-occurring phenomena are constructed, and then maps of evolving scientific and technological fields are generated using the link-node values of the networks. The major co-occurrence techniques, which include co-citation, co-word, co-nomination, co-classification, and co- authorship, will be described briefly.

Co-citation analysis is based on the principle that when a paper X1 cites two earlier papers A and B, these latter papers are 'co-cited'. The strength of such a co-citation link is determined by the number of citing papers (X1, X2, X3...) each with the pair (A, B) in their lists of cited papers (references). Paper B can also form a co-citation pair with a third paper C, etc. The next analytical step generates a structure in interlinked co-cited pairs creating sets of (co-)cited papers, that is, clustering of co-cited papers yielding aggregates of the size of scientific fields of disciplines.

Finally, the network of co-citation linkages within and between clusters is visualized through a variety of mapping techniques [Tijssen, 1994; Small, 1998].

Co-citation clusters represent research-front specialties, in terms of related scientific work. These clusters may reflect cognitive as well as social networks. However, drawbacks of co- citation analysis include: dependence on a restricted literature (cited publications); its time-lag inherent to citation data in general; the cluster algorithm may involve somewhat arbitrary setting of threshold values, yielding less informative results in case of non-optimal thresholds; and inadequate coverage of technology-related research.

Co-word has been utilized to map the evolution of science under European (mainly French) government support. It is the genesis of DT, and will be described in more detail after this brief summary of co-occurrence techniques.

Co-nomination is a particular example of the more general social network analysis used to study communication among workers in the fields of science and technology, and is not nominally a computer-based approach in the sense of the other co-occurrence techniques. Generally, in co-nomination, experts in a given field are asked to identify other experts, and then a network is generated which shows the different linkages (and the strengths of these linkages) among all the experts (and possibly their organizations and technical disciplines) identified [Georghiou, 1988]. A survey [Shrum, 1988] of the development of social network analysis traces studies in this area back at least three decades [Libbey, 1967; Blau, 1978]

Co-nomination was developed to circumvent co-citation's dependence upon databases consisting only of refereed scientific publications. It is a more direct approach of obtaining links among researchers and, if combined with other network approaches which include both links between technical fields and the link strengths [Kostoff, 1994a], could potentially incorporate links among researchers and technical fields.

Co-classification analysis operates on the co-occurrence of terms (or codes) which are used to classify publications for ease of access in bibliographic databases [Tijssen, 1994]. These indexer-given information items are derived from a thesaurus and may represent scientific (or technological) topics, specialties, or fields. Compared to key-words, subject classification terms have a well-defined and consistent meaning over the entire knowledge domain, which makes them particularly attractive for studying and depicting the main cognitive structure across large scientific and technological areas. The main practical restrictions are the fixed classification scheme, and the restricted databases which have such classifications. Moreover, classification codes are assigned primarily for information retrieval purposes and do not necessarily reflect intellectual concepts. Co-classification has been used to map the structure of chemical engineering [Van Raan, 1989] and to depict the structure of relations among all technological fields [Engelsman, 1991].

Co-authorship is the analysis of author attributes of multiple-authored papers. Co-authorship has been used to assess collaboration among individuals, and institutional co-authorship has been used to assess collaboration among institutions, through construction of scientific networks [Melin, 1996].

2.5 Co-Word Analysis

2.5.1 Origins and Overview

Co-word analysis examines phrase co-occurrence patterns in different domains, and can be traced back six decades to pioneering work in lexicography and linguistics [Hornby, 1942; De Saussure, 1949]. A summary of co-word origins, and evolution of co-word into computational linguistics, can be found in Kostoff [1993b].

Co-word analysis, in conjunction with expert interpretation, is used for two different aspects of DT. It is used iteratively to develop the final query for extracting relevant records from the source database (information retrieval), and it is used to define and inter-relate the pervasive technical themes of the technical domain being studied (research evaluation). The following two sections summarily describe the evolution of the use of co-word analysis for information retrieval and research evaluation.

2.5.2 Co-word Analysis for Information Retrieval

Many modern information retrieval techniques are iterative. They start with a group of documents known to be highly relevant to the user. These techniques extract patterns characteristic of the relevant documents and, through iterative query modification, search for other documents in the database which have similar characteristic patterns (relevance feedback) [Salton, 1990]. How the query is then expanded in practice depends on the purpose of the larger study in which the query is embedded. If the study's purpose is comprehensive coverage and analysis of a field of knowledge, then the query's expansion will be limited to characteristics of the specific field. If, however, the study's purpose is innovation and discovery of new knowledge through the process of related literatures [Swanson, 1986; Kostoff, 1999], then the query must be expanded well beyond the initial technical field, in a controlled manner.

One generic research avenue in characteristic pattern matching techniques has involved term co-occurrence embedded in a relevance feedback structure with query expansion. Term co-occurrence in information retrieval can be used to expand on an initial query, and the additional query terms allow the retrieval of relevant documents that would not have been retrieved with the initial query. These additional terms could also be used to remove irrelevant documents.

Studies related to the use of term co-occurrence in information retrieval can be traced back to at least the 1960s [Maron, 1960; Stiles; 1961; Doyle, 1962; Lesk, 1969]. These early experiments demonstrated the potential of term co-occurrence data for the identification of search term variants. The use of term co-occurrence in information retrieval was motivated initially by the heuristic observations that searchers and authors tend to use different terms to describe the same information, and consequently a number of related query terms are required for increased search effectiveness and efficiency. These observations were confirmed later by experiments [Furnas, 1987; Gomez, 1990].

Later work on query expansion related to term co-occurrence was based on probabilistic models of the retrieval process, and tried to relax some of the strong assumptions of term statistical independence that normally need to be invoked if probabilistic retrieval models are to be used [Croft, 1979; Robertson, 1976; Smeaton, 1983].

In the 90s, work in the use of term co-occurrence for query expansion has exploited large presently-available computing power to generate thesauri automatically [Crouch, 1990; Rasmussen, 1992]. A series of late 90s reported studies has used automatic indexing and co-occurrence analysis, performed on parallel computers, to generate a domain-specific thesaurus automatically [Chen, 1997; Chen, 1998].

Relevance feedback implementations were initially designed for queries and documents in vector form; i.e., query statements consisting of sets of possibly weighted search terms used without Boolean operators [Rocchio, 1971; Salton, 1971] . Since the 1980s, relevance feedback methods have been applied also to Boolean query formulations, where the process incorporates term conjuncts (derived from previously retrieved relevant documents) into revised query formulations [e.g., Salton, 1985] In the mid-90s, relevance feedback approaches with probabilistic information retrieval based on document components have been incorporated into artificial neural networks [e.g., Kwok, 1995], and have resulted in improved performance.

Another important recent study, which relates to the use of relevance feedback in DT, focused on determining the retrieval effectiveness of search terms identified by users and intermediaries from retrieved items during term relevance feedback. Terms selected from particular database fields of retrieved items during term relevance feedback were more effective than search terms from the intermediary, database thesauri or users' domain knowledge during the interaction [Spink, 1995].

Also recently, use of local context analysis [Xu, 1996], which combines global analysis [Jing, 1994; Callan, 1995] and local feedback [Attar, 1977], has generated effective information retrieval results. In this combined approach, noun groups are used as concepts and concepts are selected based on co-occurrence with query terms. Concepts are chosen from the top-ranked documents, similar to local feedback, but the best passages are used instead of whole documents. An algorithm is used to rank the concepts, and the query is then expanded.

The information retrieval approach used for the present study is based on term co-occurrence with relevance feedback for query expansion. It employs the DT algorithms, and can be used by a wide variety of analysts with little training to provide highly efficient retrieval. The mechanics are summarized in section 3, and the details are presented in Kostoff [1997a].

2.5.3 Co-word Analysis for Research Evaluation

Modern development of co-word analysis for purposes of evaluating research originated in the mid-1970s [Callon, 1979, 1983, 1986]. The method developed initially by Callon focused on analyzing the content of articles and reports. In one of the first descriptions and applications of the method [Callon, 1979], the impact of French government intervention in the field of macromolecular chemistry was examined. A database of over 4,000 articles covering the field of interest was generated. Key or index words were assigned to each article in the database. A basic assumption was that the key words describing an article had some linkages in the author's mind, and the different fields or functions represented by these words had some relation.

Each time a pair of words occurred together in the key word list of an article, it was counted as a co-occurrence of the pair. The number of co-occurrences for each pair was calculated for all the articles in the database. A co-occurrence matrix was constructed whose axes were the index words

in the database and whose elements were the number of pair co-occurrences of the index words. A two-dimensional map was constructed which would display visually the positions of the key words relative to each other based on their co-occurrence values from the matrix. While different maps had different axes pairs, the central features of the maps appeared to be display of the relationship structures, and the strength of the relationships, between the words.

There were at least two major problems with this approach: the text was not analyzed directly, and the analysis was performed on the key words. The bias and error introduced from key word analysis was unknown, but use of key word indexing continued to affect the credibility of the technique for years [Bates, 1986; Healey,1986; Leydesdorff, 1987].

Subsequent co-word studies focused on: biotechnology [Rip, 1984]; aquaculture [Bauin, 1986]; patents [Callon, 1986]; industrial ceramics [Turner, 1988]; polymer science [Callon, 1991]; neural networks [(Van Raan, l991]; chemical engineering [Peters, 1991]; combined word frequency analysis of citing articles with co-citation analysis [Braam,1991a; Braam,1991b]; and material science [Van Raan, 1996]. All of these reported studies used key words or index words, not full text.

2.6 Development of DT

Callon's classical co-word analysis does not allow the richness of the semantic relationships in full text to be exploited, and it is restricted to formally published papers. In order to allow any form of free text to be used, DT was developed.

In 1990-1991, experiments were performed at the Office of Naval Research [Kostoff, 1991] which showed that the frequency with which phrases appeared in full text narrative technical documents was related to the main themes of the text. The phrases with the highest frequencies of appearance represented the main, 'pervasive' themes of the text. In addition, the experiments showed that the physical proximity of the phrases was related to the thematic proximity. These experiments formed the basis of DT.

The DT method in its entirety requires generically three distinct steps. The first step is identification of the main themes of the text being analyzed. The second step is determination of the quantitative and qualitative relationships among the main themes and their secondary themes. The final step is tracking the evolution of these themes and their relationships through time. The first two steps will be summarized now. Time evolutions of themes have not yet been performed. First, the frequencies of appearance in the total text of all single word phrases (e.g., Matrix), adjacent double word phrases (e.g., Metal Matrix), and adjacent triple word phrases (e.g., Metal Matrix Composites) are computed. The highest frequency significant technical content phrases are selected by topical experts as the pervasive themes of the full database.

Second, for each theme phrase, the frequencies of phrases within +/- M (nominally 50) words of the theme phrase for every occurrence in the full text are computed, and a phrase frequency dictionary is constructed. This dictionary contains the phrases closely related to the theme phrase. Numerical indices are employed to quantify the strength of this relationship. Both quantitative and qualitative analyses are performed by the topical expert for each dictionary (hereafter called cluster) yielding, among many results, those subthemes closely related to and supportive of the main cluster theme.

Third, threshold values are assigned to the numerical indices, and these indices are used to filter out the most closely related phrases to the cluster theme. However, because numbers are limited in their ability to portray the conceptual relationships among themes and sub-themes, the qualitative analyses of the extracted data by the topical experts have been at least as important as the quantitative analyses. The richness and detail of the extracted data in the full text analysis allows an understanding of the theme interrelationships not heretofore possible with previous text abstraction techniques (using index words, key words, etc.).

At this point, a variety of different analyses can be performed. For databases of non-journal technical articles [Kostoff, 1992, 1993a], the final results have been identification of the pervasive technical themes of the database, the relationship among these themes, and the relationship of supporting sub-thrust areas (both high and low frequency) to the high-frequency themes. For the more recent studies in which the databases are journal article abstracts and associated bibliometric information (authors, journals, addresses, etc), the final results have also included relationships among the technical themes and authors, journals, institutions, etc [e.g., Kostoff, 1998a].

The study reported in the present paper is in the latter (journal article abstract) category. It differs from the most recently published paper in this category [Kostoff, 1998a] in four significant respects. First, the topical domain is completely different (hypersonic and supersonic flow over aerodynamic bodies vs. utilization of near-earth space). Also, the present topic is focused on a single technology (hypersonic/ supersonic flow) vs. the assemblage of technical disciplines which constitute the study of near-earth space. Second, normalized distribution functions of completed DT studies were developed for cross-domain comparisons, and conclusions were drawn about these distributions which appeared to transcend domains. Third, a taxonomy of database megathemes was developed by parametrically tracking the clustering formation of these categories. Fourth, a comparison of characteristics of highly-cited and poorly-cited hypersonics papers was presented. The results offered a number of interesting insights, and led to a subsequent comparison of non-hypersonics papers of American and Russian authors (not reported here).

2.7 Evolution of DT into Textual Data Mining

Recent evaluations of real-world textual Data Mining applications (unpublished) across a number of organizations showed a strong decoupling of the research performer from the Data Mining user. The performer tended to focus on the development of exotic automated techniques, to the relative exclusion of the components of judgement necessary for user credibility and acceptance. Consequently, the Data Mining techniques actually employed by most of the potential users examined centered on reading of copious numbers of articles obtained by the simplest of queries. The DT process reported in this paper represents the framework of a Data Mining approach which will couple the Data Mining research and associated computer technology processes much more closely with the Data Mining user. Strategic database maps will be developed on the front end of the process using bibliometrics and DT, with heavy involvement from topical domain experts (either users or their proxies) in the DT component of strategic map generation. The strategic maps themselves will then be used as guidelines for detailed expert analysis of segments of the total database. The authors believe that this is the proper use of automated techniques for Data Mining: to augment and amplify the capabilities of the technical expert by providing insights to the database structure and contents, not to replace the technical domain experts by a combination of machines and non-experts.

3. DATABASE GENERATION

Now the present study methods and results will be described. The key step in the HSF literature analysis is the generation of the database. For the present study, the database consists of selected journal and conference proceeding records (including authors, titles, journals, author addresses, author keywords, abstract narratives, and references cited for each paper) obtained by searching the CD-ROM version of the Science Citation Index (SCI), and the Engineering Compendex (EC) for HSF articles. This version of the SCI accesses about 3200 journals (mainly in physical and life sciences basic research) and the EC accesses about 2600 journals and conference proceedings (mainly in applied research and technology).

This database selected represents a fraction of the available HSF literature. It does not include the large body of classified literature, or company proprietary technology literature. It does not include the large body of technical reports on HSF. It covers a finite slice of time (1993 to mid-1996). The database used represents the bulk of the peer-reviewed high quality HSF science and technology, and is a representative sample of all HSF science and technology in recent times.

To extract the relevant articles from the SCI and EC, the title, keyword, and abstract fields were searched using keywords relevant to HSF, although different procedures were used to search the title and abstract fields [Kostoff, 1997a]. The resultant abstracts were culled to those relevant to HSF. The search was performed with the aid of two powerful DT tools (multi-word phrase frequency analysis and phrase proximity analysis) using the process of Simulated Nucleation [Kostoff, 1997a].

An initial query of HYPERSONIC* OR SUPERSONIC* produced two groups of papers: one group was judged to be relevant to the subject matter by a domain expert, the other was judged to be non- relevant. Gradations of relevancy or non-relevancy were not considered. An initial database of titles, keywords, and abstracts was created for each of the two groups of papers. Phrase frequency and proximity analyses were performed on this textual database for each group. The high frequency single, double, and triple word phrases characteristic of the relevant group, and their boolean combinations, were then added to the query to expand the papers retrieved. Similar phrases characteristic of the non-relevant group were added to the query (to the NOT boolean) to contract the papers retrieved. The process was repeated on the new database of titles, keywords, and abstracts obtained from the search. A few more iterations were performed until the number of records retrieved stabilized (convergence).

The authors believe that the 'purity' and completeness of the database of topically relevant records obtained using Simulated Nucleation is a key reason that the invariance of most of the normalized bibliometric distributions across different topical domains can be displayed (see sections 4.1 and 4.2 for the normalized bibliometric distribution functions). One beneficial value of utilizing Simulated Nucleation is that the search terms are obtained from the words of the authors in the SCI and EC databases, not by guessing on the part of the searcher.

4. RESULTS

First, the results of the bibliometric analyses will be presented, then followed by the results of the DT analyses. The SCI and EC bibliometric fields incorporated into the database included, for each paper, the author, journal, institution, and keywords. In addition, the SCI included references for each paper. Due to space limitations, the integrated results from analysis of each of these fields will now be presented for the SCI database only. As contrasted with the near-earth space [Kostoff, 1998a] analyses results, where the SCI and EC papers were obtained from essentially different journals, the HSF SCI and EC journals accessed overlapped strongly. It may be that the journals in which HSF papers are published are not as strongly stratified into basic and applied research as are the near-earth space journals, but rather are more inclusive of different development categories.

4.1 Most Published Authors, Journals, Organizations, Countries

The first group of metrics presented is counts of papers published by different entities. These metrics can be viewed as output and productivity measures. They are not direct measures of research quality, although there is some threshold quality level inferred due to their publication in the (typically) high caliber of journals accessed by the SCI.

4.1.1 Prolific Authors

The author field was separated from the database, and a frequency count of author appearances was made. In the SCI database results, there were 2483 different authors, and 3372 author listings (the occurrence of each author's name on a paper is defined as an author listing). While the average number of listings per author is about 1.38, the most prolific authors (e.g., PILYUGIN, MCDANIEL, BOYD, GRASSO) have listings about an order of magnitude greater than the average. There were 1284 papers retrieved, yielding an average of 2.63 authors per paper. Previous studies of the technical fields of near-earth space [Kostoff, 1998a] and of Chemistry [Kostoff, 1997b] as represented by the Journal of the American Chemical Society yielded 3.37 authors per paper for the space results and three authors per paper for the Chemistry results. A previous study on the non-technical field of research impact assessment (RIA) yielded about 1.3 authors per paper. See Table 1 for summary statistics of pre-2000 studies.

TABLE 1 - DT STUDIES OF TOPICAL FIELDS

TOPICAL AREA	NUMBER OF SCI ARTICLES	YEARS COVERED
CHEMISTRY (JACS)	2150	1994
NEAR-EARTH SPACE (NES)	5480	1993-MID 1996
HYPERSONICS (HSF)	1284	1993-MID 1996

RESEARCH ASSESSMENT (RIA)	2300	1991-EARLY 1995
FULLERENES (FUL)	10515	1991-MID 1998
AIRCRAFT (AIR)	4346	1991-MID 1998
HYDRODYNAMICS (HYD)	4608	1991-MID 1998

In three of the previous studies, it was concluded that the RIA projects tended to be individual efforts by their intrinsic nature, while the space and Chemistry projects tended to involve large teams in many of the experimental projects. One would expect that the hypersonic/ supersonic-related projects/ papers from the present study would reflect large collaborative groups. Especially in the wind-tunnel and flight experiments, large facilities, efforts, and costs are involved, and typically many different experiments are performed. Many of the efforts tend to involve multiple disciplines as well.

The presence of a moderate number of collaborators per hypersonic paper means that these large experimental hypersonic research projects do not dominate to the extent expected, and that individual small-scale projects play an important role in hypersonics research. Later results from the keyword phrase frequency analyses and other phrase frequency results seem to support this conclusion, and substantiate the picture of much hypersonics research as smaller efforts in Computational Fluid Dynamics. It should be re-emphasized that these conclusions apply to hypersonics research as published in the open literature. The conclusions could change for technology development, where larger team efforts would probably be the norm, and for classified or proprietary research.

FIGURE 1

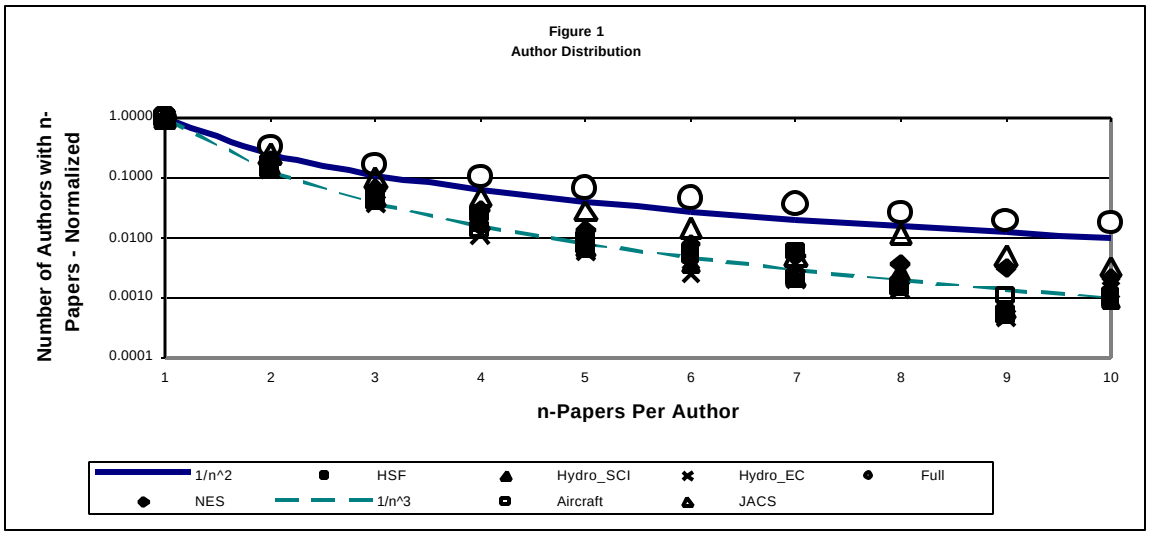

Figure 1 shows the distribution function of SCI author listing frequency for the NES, JACS, HSF, AIR, HYD, and FUL databases. The abcissa is the number of author listings n, and the ordinate is the number of authors who have author listing n. In each case, the distribution function has been normalized to the number of authors who have one listing in the respective databases. The graph is plotted on a semi-log scale to stretch the lower ordinate region.

The solid line on Figure 1 is the nominal (1/n^2) Lotka's Law (Lotka, 1926) distribution. With the exception of the FUL data, all of the experimental data decline much steeper than the (1/n^2) Law predicts, centering about a (1/n^3) distribution. One interpretation of this observation is that Lotka concentrated on only the very core journals in the disciplines studied. These journals tend to accept relatively more contributions from the prolific and recognized researchers than the non-core journals.

4.1.2 Journals Containing Most HSF Papers

A similar process was used to develop a frequency count of journal appearances. In the SCI database, there were 277 different journals represented, with the median journal containing one paper, and an average of 5.7 papers per journal. The journals containing the most HSF papers (e.g., AIAA JOURNAL, JOURNAL OF SPACECRAFT AND ROCKETS, JOURNAL OF PROPULSION AND POWER) had an order of magnitude more papers than the average. Bradford's law [Bradford, 1934] for journal publications can be re-stated as: if the journals for a bibliography are grouped in order of decreasing publications, such that each group of journals contains the same number of papers, then the ratio of number of journals in each successive group will be a constant greater than unity. For the HSF database, the first group selected contains one journal with 231 papers (AIAA JOURNAL); the second group has 3 journals with 237 papers; third group 9 journals with 229 papers; fourth group 25 journals with 229 papers; and fifth group 70 journals with 229 papers. The ratio of numbers of journals per group between successive groups is approximately three, in excellent agreement with Bradford's law.

FIGURE 2

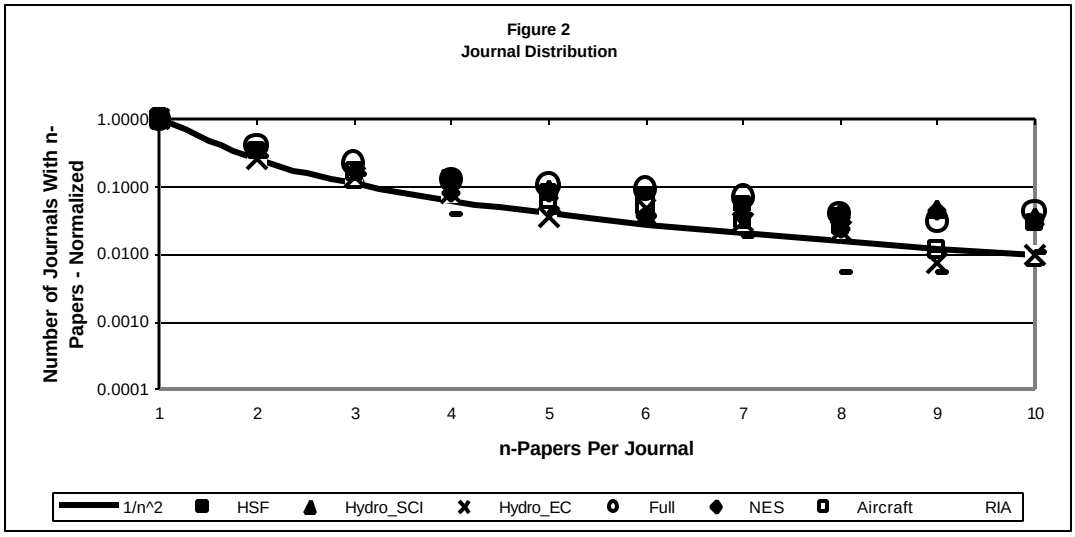

Figure 2 shows the distribution function of SCI journal frequency for the FUL, AIR, HYD, HSF, NES, and RIA databases. The JACS database was derived from one journal only, The Journal of the American Chemical Society, and therefore was not applicable to this chart. The abcissa is the number of papers n from the relevant database published in a given journal, and the ordinate is the number of journals which contain n papers. In each case, the distribution function has been normalized to the number of journals that contain one relevant paper. Again, because of the strong initial gradients, the graph is plotted on a semi-log scale.

The solid line in Figure 2 is a (1/n^2) distribution, and represents a lower bound of all the experimental data.

4.1.3 Institutions Producing Most HSF Papers

A similar process was used to develop a frequency count of institutional address appearances. There were 661 different organizations listed in the SCI author address organizations, with the median organization producing one paper, and an average of 1.94 papers per organization. The institutions producing the most HSF papers (e.g., NASA, RUSSIAN-ACAD-SCI, OFF-NATL-ETUD-&-RECH-AEROSP, MOSCOW-MV-LOMONOSOV-STATE-UNIV) were more than an order of magnitude more productive than the average. The NASA labs are by far the most productive of any of the institutions in terms of papers published, although no statements can be made about their production efficiency, since research expenditures were not included in this study. It should be noted that many different organizational components may be included under the single organizational heading (e.g., Harvard Univ could include the Chemistry Department, Biology

Department, Physics Department, etc.). Lack of space precluded printing out the components under the organizational heading.

FIGURE 3

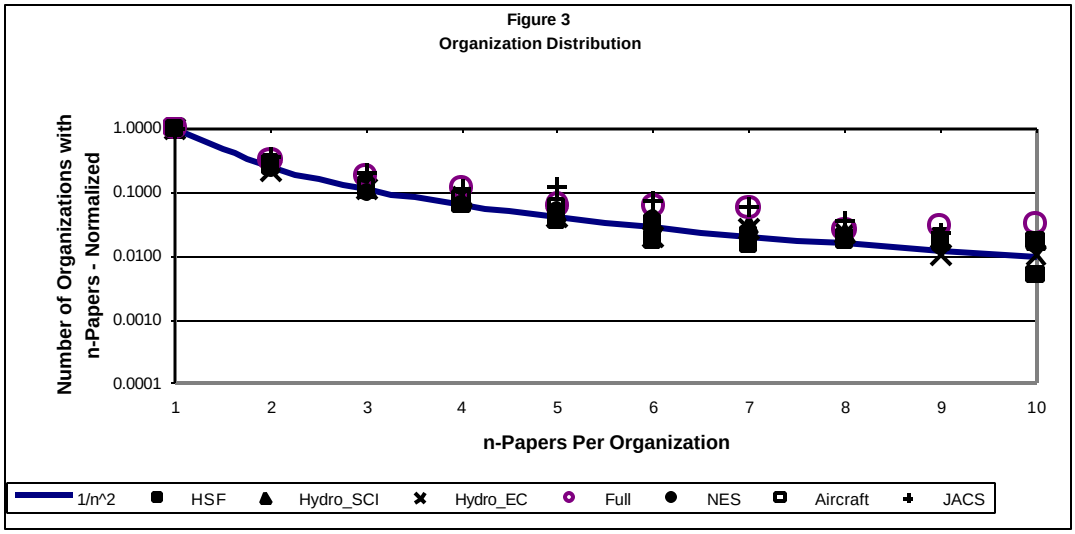

Figure 3 shows the distribution function of SCI institution frequency for the HSF, NES, JACS, AIR, HYD, and FUL databases. The abcissa is the number of papers n in the database produced by a given institution, and the ordinate is the number of institutions that produced n relevant papers. In each case, the distribution function has been normalized to the number of institutions that produced one relevant paper.

The data center around a (1/n^2) distribution remarkably well. For a (1/n^2) distribution, the number of organizations that generate three papers is about eleven percent of the organizations that generate one paper only. Also, integrating this distribution function shows that more than 67% of the papers result from organizations that produce three or less papers.

4.1.4 Countries Producing Most HSF Papers

There were 53 different countries listed in the SCI results. The dominance of a handful of countries was clearly evident. The UNITED STATES is about an order of magnitude more prolific than its nearest competitor (RUSSIA), and is as prolific as its major competitors combined (RUSSIA, JAPAN, FRANCE, UK, GERMANY, ITALY, CANADA). In the four initial studies performed using the present approach (Research Impact Assessment [RIA], Chemistry [JACS], Near- Earth

Space, Hypersonic-Supersonic Flow), this dominant relationship between the United States and its nearest competitors was observed. Generically, the western democracies tend to be the most prolific. In addition, Japan is in the first JACS tier and second RIA tier; Hungary is high in RIA; and India and Russia are both well into the second RIA and JACS tiers. A 1997 study [Anwar, 1997] listed the papers contributed by the top 50 nations to the world science literature; i.e., numbers of publications in the SCI. The top performers are in line with the bibliometric results of the four DT studies.

4.2 Most Cited Authors, Papers, Years, and Journals

The second group of metrics presented are counts of citations to papers published by different entities. While citations are ordinarily used as impact or quality metrics, much caution needs to be exercised in their frequency count interpretation, since there are numerous reasons why authors cite or do not cite particular papers [Kostoff, 1997c, 1998b; MacRoberts, 1996].

4.2.1 Most Cited Authors

The citations in all 1284 SCI papers were aggregated into a file of 26768 entries, yielding an average of 20.9 references per paper. There were 11138 different authors cited, with an average of 2.4 citations per author. A relatively few percent received large numbers of citations (e.g., TAM, PAPAMOSCHOU, ROE, PARK, ANDERSON, BILLIG). However, the most cited authors, while prolific, are not the most prolific authors, and vice versa. For example, the three most highly cited authors (TAM, PAPAMOSCHOU, ROE) ranked numbers 35, 45, and off the chart, respectively, in the prolific authors list. The three most prolific authors (PILYUGIN, MCDANIEL, BOYD) all ranked number 53 (tied) in citability.

Part of this difference may be due to the time lag between the highly cited authors' productivity at the time their highly cited papers were written and their productivity today, as well as the phase in their career of the prolific authors. Another partial explanation may be the intrinsic nature of the papers; the large numbers of papers produced may reflect more applied papers, which lend themselves more to shorter-term production line type output. Stated differently, the time required to produce a fundamental seminal highly cited paper probably does not allow overly high volumes of papers to be produced.

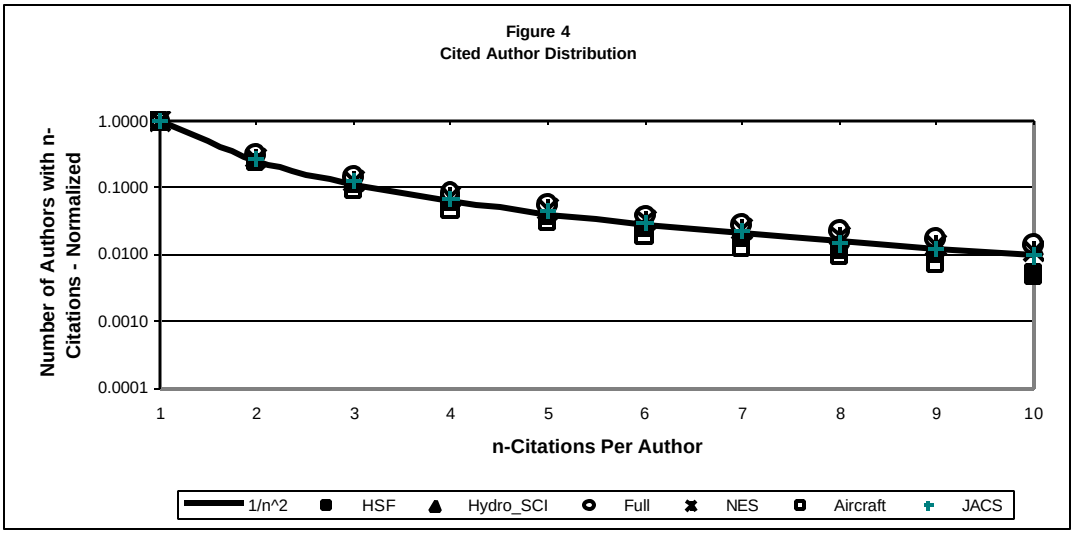

FIGURE 4

Figure 4 shows the distribution function of author citation frequency for the FUL, NES, HSF, JACS, AIR, and HYD databases. The abcissa is the total number of citations n received by a given author, and the ordinate is the number of authors that received n total citations. In each case, the distribution function has been normalized to the number of authors that received one citation.

The data cluster very closely around a (1/n^2) distribution, making this distribution far more universal than the somewhat discipline-dependent author publishing distribution.

Integration of this distribution function shows that over 67% of the citations are from authors cited three times or less. Section 4.2.2.2 describes a focused study which examined all citations from a small sample group of hypersonic papers. By far, the largest number of papers (the mode) in that study received no citations. The emerging picture is that HSF has a relatively low level of activity due to the low magnitude of citations, and that most authors in HSF have no cites or very few cites. Only a small fraction of the papers are seminal documents which are cited highly (in relative, not absolute, terms).

Some caveats are in order at this point. The citation data for Figures 4, 5, 6 represents citations generated only by the 1284 HSF papers. It does not represent all the citations received by the references in those 1284 papers; these references in the 1284 HSF papers could have been cited additionally by papers in other technical disciplines. In addition, since very recent papers are included in the references, there is probably some skewing of the distribution function toward lower numbers of citations in these figures relative to distribution functions which don't include very recently published references. Recent papers don't have sufficient time to accumulate more than a small number of citations.

Conversely, the sample studies in section 4.2.2.2 do not have the two limitations described in the above paragraph. In the sample study, a small number of papers was selected. All citations to those papers from SCI papers in all fields were included, and a 3-4 year time interval between date of publication and the present was chosen to allow reasonable numbers of citations to accumulate.

4.2.2 Most Cited Papers

4.2.2.1 Aggregate Distribution Functions

There were 20950 different papers cited, with an average of 1.27 citations per cited paper. Relatively few papers were highly cited (e.g., ROE, 1981; PAPAMOSCHOU, 1988; ANDERSON, 1989). From the citation year results, the most recent papers are the most highly cited. This reflects a rapidly evolving field of research.

FIGURE 5

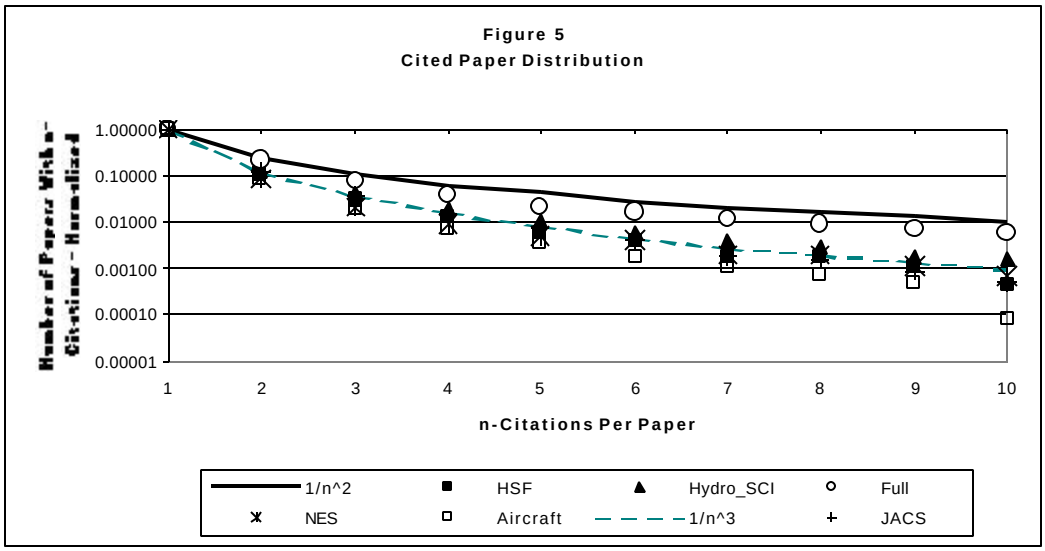
Figure 5
Cited Paper Distribution
1.00000
0.10000
0.01000
0.00100
0.00010
0.00001
1
2
3
4
5
6
7
8
9
10
n-Citations Per Paper
1/n^2
HSF
Hydro_SCI
Full
NES
Aircraft
1/n^3
JACS

Figure 5 shows the distribution function of paper citation frequency for the NES, JACS, HSF, AIR, HYD, and FUL databases. The abcissa is the total number of citations n received by a given paper, and the ordinate is the number of papers that received n total citations. In each case, the distribution function has been normalized to the number of papers that received one citation.

For five of the six topical fields presented, the data follow a (1/n^3) distribution very closely, as contrasted with the (1/n^2) distribution for author citations. Examination of the five topical studies that produced the five sets of data showed that each of the highly cited authors had a wide range of citations for his/ her different papers. For any given highly cited author, most papers will receive few citations. It is the infusion of numbers of lowly cited papers from the highly cited authors that expands the pool of lowly cited papers in Figure 5, and results in the conversion of the (1/n^2) distribution of Figure 4 to the (1/n^3) distribution of Figure 5. This effect appears to transcend the five different science and technology topical fields, and to be almost universal based on the limited data presented for the six topical science and technology fields. This relation, the Kostoff-Eberhart-Toothman (KET) Law [11], can be re-stated as follows: for a topical science and technology field, the ratio of the normalized number of authors with n citations per author to the normalized number of papers with n citations per paper is n, for low to moderate values of n.

4.2.2.2 Characteristics of Highly-Cited and Poorly-Cited Papers

To ascertain whether any relationship between highly cited and lowly cited papers and their associated journals and performing organizations could be observed, the characteristics of samples of highly cited and lowly cited papers were analyzed. The database used to extract the samples was the expanded Web version of the SCI. In contrast to the CD-ROM version of the SCI used to obtain the bulk data for this paper, the Web version has 60% more journals (~5200), and is more convenient for performing citation analyses.

All records in the Web version which contained the term HYPERSONIC (a small subset of the supersonic/ hypersonic field) and were published in 1993 were examined. Analysis of the 140 applicable records showed that the journals and organizations in the highly cited sample were, on average, the very top echelon of the total database journals and organizations in terms of numbers of papers published and contributed, respectively.

The journals with a high ratio of highly cited papers to zero cited papers tend to emphasize the more fundamental research. The journals with a low ratio of highly cited papers to zero cited papers tend to emphasize the more applied research. The fact that the applied papers are being cited less than the more fundamental papers does not mean they are less useful or of lower quality; they may be of substantial use to developers, who publish much less than researchers, and this more practical use would not be reflected in the present type of bibliometrics study.

In summary, this small sample analysis led to the following conclusions for hypersonic flow. Fundamental research papers are more likely to be cited than applied research papers; university papers are more likely to be cited than industry papers; the journals which contain concentrations of highly cited papers are also the core journals in terms of papers published; NASA produced many papers (147 in the total CD-ROM database), and had a substantial fraction of the highly cited

papers; Russia produced slightly more papers than NASA (169 in the total CD-ROM database), and had almost no highly cited papers.

The NASA/ Russia citation differential led to another short study which examined American/ Russian differentials in HSF citations. Two groups of papers were generated. The first group consisted of all papers (from the web version of the SCI) published in 1993/ 1994 by the three most prolific HSF Russian authors identified in the main hypersonic study; the second group included all papers by the three most prolific HSF American authors identified in the main hypersonic study. There were 12 papers in the first (Russian) group, and 36 papers in the second group.

The average cites per Russian paper was 0.4, and all these cites were self-cites (There is nothing intrinsically wrong with self cites; in those cases where the author has done the pioneering work in the field, self-cites are most appropriate. However, when all cites are self-cites, then the true impact of the paper on the larger scientific community must be called into question).

The average cites per American paper was three, half of which were self-cites. While all these citation numbers reported are quite small, reflecting the low level of reported research effort in this technical field, there is obviously a systemic difference between the citations received by the Russian and American papers.

There are two crucial pieces of data missing from these two short studies (and from most bibliometrics analyses) which prevent harder conclusions about quality and value to be drawn. The amount of research effort represented by each paper, and the eventual use of the results from each paper, are unknown to the analyst. Thus, the number of highly cited papers per dollar of research investment (or some similar research efficiency metric), probably a better measure of value than pure numbers of papers or highly cited papers, cannot be stated. Also, the quality of the eventual hypersonic vehicles which resulted from the papers' research, probably a better real-world quality measure than numbers of cited papers, was not tracked and cannot be stated. In addition, the papers in these two short studies were not read in detail independently by hypersonic flow experts, and thus their quality could not be gauged independently from another perspective and correlated to the citation results.

4.2.3 Most Cited Journals

There were 9498 different journals and other sources cited. Relatively few sources were highly cited (e.g., AIAA JOURNAL, JOURNAL OF FLUID MECHANICS, JOURNAL OF CHEMICAL PHYSICS, ASTROPHYSICS JOURNAL, JOURNAL OF COMPUTATIONAL PHYSICS, JOURNAL OF SPACECRAFT AND ROCKETS). There is probably somewhat more correlation between journals which are highly cited and which contain large numbers of HSF papers than between highly prolific and cited authors. The time span over which a journal develops and maintains a reputation for high quality is long compared to the gap between publication and citation, and one should expect that in the steady state the journals which publish many HSF papers would also publish the higher quality papers. To the degree that the most highly cited papers have the highest quality, the voluminous content journals should contain a larger share of the higher cited papers.

AIAA JOURNAL tended to publish many HSF papers and be highly cited, but JOURNAL OF CHEMICAL PHYSICS was highly cited and did not appear to publish many HSF papers under the present definition. Analogous to the results of the small sample study presented in section 4.2.2.2, one possible explanation is that the HSF published papers are slightly more applied than some of their references. If JOURNAL OF CHEMICAL PHYSICS tends to contain very fundamental papers typically, it would serve mainly as a citing source for HSF papers, but not a publishing source for HSF papers. The more fundamental journals (ASTROPHYSICS JOURNAL, JOURNAL OF COMPUTATIONAL PHYSICS, PHYSICS OF FLUIDS, CHEMICAL PHYSICS LETTERS, JOURNAL OF PHYSICAL CHEMISTRY) rank higher on citations relative to their publication rankings, while the more applied journals (JOURNAL OF PROPULSION AND POWER, HIGH TEMPERATURE, JOURNAL OF THERMOPHYSICS AND HEAT TRANSFER) rank higher on publications relative to their citation rankings.

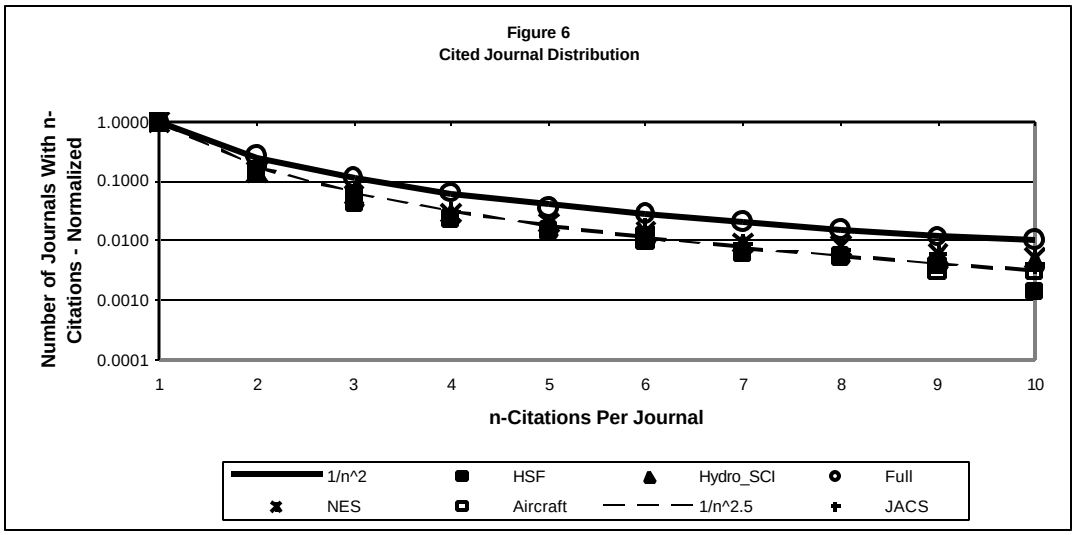

FIGURE 6

Figure 6 shows the distribution function of journal citation frequency for the NES, JACS, HSF, AIR, HYD, and FUL databases. The abcissa is the total number of citations n received by a given journal, and the ordinate is the number of journals that received n total citations. In each case, the distribution function has been normalized to the number of journals that received one citation. The data follow approximately a (1/n^2.5) distribution.

There are some important implications to be drawn from these journal distribution functions and tabulated metrics with regard to text mining, and these conclusions will be addressed briefly. During the development of the Bradford's Law metric mentioned previously, the number of journals in successive iso-paper groups was computed. In addition, the number of journals in successive iso-citation groups was computed for NES, HSF, and AIR, to ascertain whether a Bradford's Law for citations was operable. The ratio between iso-citation groups was less regular than the ratio between iso-paper groups, and seemed to vary between 1.5 and 2 for the three studies.

However, a very important message can be extracted from this data, namely, that a potential substantial capability increase (for an organization involved in S&T) from a successful text mining program is possible. Consider the aircraft results as an example (while actual numbers may differ among disciplines, the conclusions drawn are probably applicable to any technical discipline).

There are over 700 different journals that contain aircraft-related papers. The core (first) journal group (for the Bradford's Law computation) contains three journals. There are about 6 journal groups that contain the total number of over 700 journals, the first five groups being iso-paper, and the last somewhat less (essentially, the remainder). Thus, the core journal group contains about 18-20% of the total number of papers. For a technical manager or performer to be considered a true expert in all aspects of aircraft S&T, this individual would have to be familiar with the results from the aircraft papers in most of the more than 700 journals. One would suspect that bench-level aircraft experts, such as field managers, don't read more than the first two core groups on a regular basis, and this is probably a very generous estimate. Thus, these experts may be familiar with 30-40% of the relevant literature within the focused field; they would be far less familiar with complementary disparate-discipline literatures from which novel concepts could be extrapolated to benefit aircraft S&T.

In addition, one would suspect that program managers, at the Federal level or in the field, who have broad responsibilities for aircraft S&T development (or of any technical discipline/multi-discipline development), don't have time to read much more than the main core group, if that much. Thus, they are probably familiar with 10% of the relevant literature, or less, and probably far less familiar with the disparate discipline literature.

One might argue that most of the good papers are contained in the first or second core journal groups, and all that is required for effective coverage is to read the journal papers in the first one or two groups. However, if citations are used as one measure of quality, the results show that citations are at least as widely spread out among the journals as actual publications. In fact, because the most highly cited journals are not necessarily those with the most publications, the spreading among journals may be broader than the results above suggest.

One might further argue that the previous paragraph aggregates citations over papers to draw journal citation conclusions; that the most highly cited papers are contained in the first or second core groups, and all that is required for effective coverage is to read the first one or two groups. Again, the data do not support this assertion.

The ten most highly cited papers in the aircraft study were examined. It was found that none of these ten were contained in the first core group journals, and only one of these ten was contained in the second core group. One could argue that aircraft is a very broad field, and citations would more likely be aimed at papers in focused specialty journals in the lower groups than at the broader coverage journals in the higher groups. The ten most highly cited papers in the hypersonics study were then examined; hypersonics constituted a more focused technical area. It was found that two of these ten were contained in the first core group, and four of these ten were contained in the first and second core groups. If one assumes that literature coverage should encompass the more fundamental highly cited papers/ journals as well as the more applied perhaps less cited papers/ journals, then it is important that all these types of journals be included in maintaining cognizance of the technical field of interest.

Obviously, citations are not the only measure of quality, and journal research papers accessed by the SCI are not the only source of useful literature information. Technical reports accessed by NTIS, technology papers/ conference proceedings accessed by EC, program narratives accessed by

RADIUS, and patents accessed by the patent database are other sources of useful information. The presence of these other quality measures besides citations, and the presence of other data sources, further expands the number of articles/ documents to be read to maintain currency in the quality S&T, and results in even a smaller fraction of the literature accessed by any individual.

Thus, based on the results from these three different SCI bibliometric approaches (publications, aggregate citations, highly cited papers), one can conclude that (at least for the fields examined) confining one's reading to the first one or two core journal groups will exclude many high quality documents. Text mining can make the user aware of these omitted papers in the target field, and, equally important, can make the user aware of papers in disparate disciplines that could impact the target field.

The argument could then be made that the literature is only one source of information. All the other useful sources are in fact accessed through proposals, workshops, site visits, and contacts. However, all these other sources are limiting as well. Consider workshops, for example. They contain a small fraction of the technical community; they tend to attract many repeat performers; they may or may not be representative of the community, depending on how they were selected and the size of the workshop. In most workshops, the focus is on a limited target discipline. Representatives from disparate disciplines who could impact the target discipline with innovative concepts are usually not present. The attendees tend to use the workshop, or expert panel, as a forum to sell their own approaches. Their willingness to share real cutting-edge approaches in an open forum (or any forum) is questionable. Workshops tend to be dominated by forceful personalities, adding further skewing to their results.

However, text mining could potentially support and add value to workshops and expert panels as well, and complement their strengths to provide a more comprehensive and balanced product. In conclusion, this brief discussion shows by example that text mining allows informed access to a wide body of literature not accessed presently. It demonstrates further that this non-accessed literature has high quality components and is important; therefore, its availability through text mining offers a potential new or enhanced capability to support program management.

4.3 Most Frequently Used Keywords

The frequency distributions of the keywords associated with each paper were analyzed. The EC numbers are about an order of magnitude higher than the SCI numbers, but the inclusion of required classification categories in the EC keyword listings probably accounts for a substantial part of this difference.

While there appears to be a strong overlap in the primary focus of the SCI and EC as evidenced by the keywords (MATHEMATICAL MODELS, SUPERSONIC FLOW, SHOCK WAVES, BOUNDARY LAYERS, TURBULENCE), the SCI keywords tend to emphasize the more fundamental aspects of hypersonic/ supersonic flow (INSTABILITY WAVES, NAVIER-STOKES EQUATIONS, HYPERBOLIC CONSERVATION LAWS, NUMERICAL SIMULATIONS, ASYMPTOTIC ANALYSIS, INDUCED IODINE FLUORESCENCE, MACH REFLECTION), while the EC keywords tend to emphasize the more applied aspects of hypersonic/ supersonic flow (WIND TUNNELS, SUPERSONIC AIRCRAFT, HEAT TRANSFER, AERODYNAMIC LOADS, PRESSURE EFFECTS, RAMJET ENGINES, LIFT, SPACECRAFT). The SCI

keywords also emphasize computational approaches, thereby substantiating the conclusion reached in the earlier bibliometrics author section that hypersonics research has a large individual investigator component focusing on CFD. The EC keywords are more inclusive of experimental methods.

4.4 Phrase Frequency Analysis - Pervasive Themes

High frequency single, double, and triple word phrases (from the text of the database), whose technical content was deemed by topical experts to be significant, were identified as the pervasive themes. Non-technical content phrases, trivial phrases, etc., were eliminated from the analysis. In this particular exercise, the database was split into two parts, titles and abstracts, and the analysis was done on each part. Since the highest frequency phrases from the title and abstract databases were very similar, only raw data outputs from the abstract database will be presented here.

Tables 2 and 3 contain only the highest frequency adjacent double and triple word phrases, respectively. In these tables, the number preceding the phrase is the frequency of appearance of the phrase in the database. These tabulated phrases, and other high frequency phrases not shown due to space limiations, were examined by an expert in the technical area of HSF. The high frequency phrases deemed significant by the expert on the basis of his HSF experience were contextually integrated to form the following coherent picture of the main database structural elements.

TABLE 2 - ABSTRACT DOUBLE WORD FREQUENCIES

HSF SCI/ EC

627 BOUNDARY LAYER	447 MACH NUMBER	275 SHOCK WAVE
249 MACH NUMBERS	230 EXPERIMENTAL DATA	218 HEAT TRANSFER
213 SUPERSONIC FLOW	190 NAVIER-STOKES EQUATIONS	
155 SHOCK WAVES	142 SHEAR LAYER	141 FLOW FIELD
132 REYNOLDS NUMBER	132 WIND TUNNEL	117 HYPERSONIC FLOW

CODE: THE NUMBER PRECEDING EACH WORD PAIR REPRESENTS THE NUMBER OF TIMES THE WORD PAIR APPEARED IN ALL THE ABSTRACTS OF THE LITERATURE DATABASE

TABLE 3 - ABSTRACT TRIPLE WORD FREQUENCIES

HSF SCI/ EC

98 ANGLE OF ATTACK	89 ANGLES OF ATTACK
60 TURBULENT BOUNDARY LAYER	56 SIMULATION MONTE CARLO
55 COMPUTATIONAL FLUID DYNAMICS	45 VISCOUS SHOCK LAYER
33 EQUATIONS OF MOTION	30 SUPERSONIC MIXING LAYER
27 COMPRESSIBLE NAVIER-STOKES EQUATIONS	27 CONVECTIVE MACH NUMBER
27 FINITE ELEMENT METHOD	27 FREESTREAM MACH NUMBER

From a global perspective, the SCI database portrays the major focus of HSF to be SUPERSONIC/ HYPERSONIC MACH NUMBER flows over simple shapes (FLAT PLATE LEADING EDGE) at ANGLES OF ATTACK containing BOW SHOCKS and OBLIQUE SHOCKS. The experimental focus is WIND TUNNEL TESTS with measurement emphasis on SURFACE HEAT TRANSFER and SURFACE PRESSURE DISTRIBUTIONS within the VISCOUS SHOCK LAYER, SHEAR LAYER, SUPERSONIC MIXING LAYER, and TURBULENT BOUNDARY LAYER; the analytical focus is NUMERICAL SIMULATION by COMPUTATIONAL FLUID DYNAMICS with FINITE ELEMENT and MONTE CARLO SIMULATION, using the COMPRESSIBLE NAVIER-STOKES EQUATIONS to model the VISCOUS SHOCK LAYER and near-body region, and using the EULER EQUATIONS to model the outer inviscid region. Conspicuous by their absence are exotic gas mixtures (helium, hydrogen, etc) which would simulate other planetary atmospheres, exotic body shapes which would simulate novel vehicle designs, and real gas effects (dissociation, ionization, radiation, etc) which would accompany very high Mach numbers characteristic of planetary entry speeds.

This analytical procedure, and subsequent analytical procedures based on the phrase proximity results (described later), are not independent of the analyst's domain knowledge; they are, in fact, expert-centric. The computer techniques play a strong supporting role, but they are subservient to the expert, and not vice versa. The computer-derived results help guide and structure the expert's analytical processes; the computer output provides a framework upon which the expert can construct a comprehensive story. The conclusions, however, will reflect the biases and limitations of the expert(s).

For example, the expert used in the present study had experience in very high speed hypersonic flow (typically Mach Number >20) from the space program. Consequently, the analytical perspective and especially the perceived literature gaps (no exotic gas mixtures characteristic of planetary atmospheres, no high temperature dissociative and radiative phenomena) were reflective of high-speed phenomena, and might not have been easily identified by an expert with the lower speed lower temperature military hypersonic flow experience (typically Mach Number ~6-8, in terrestrial atmospheres). Conversely, a military hypersonics expert might have readily perceived gaps not immediately identifiable by the space hypersonics expert. Thus, a fully credible analysis requires not only domain knowledge by the analyst(s), but probably domain knowledge representing a diversity of backgrounds.

4.5 Phrase Proximity Analysis - Relationships Among Themes

4.5.1. Background

To obtain the theme and subtheme relationships, a phrase proximity analysis is performed about each theme phrase. Typically, forty to sixty multi-word phrase themes are selected from a multi-word phrase analysis of the type shown above. For each theme phrase, the frequencies of phrases within +-50 words of the theme phrase for every occurrence in the full text are computed. A phrase frequency dictionary is constructed which shows the phrases closely related to the theme phrase. Numerical indices are employed to quantify the strength of this relationship. Both quantitative and qualitative analyses of each phrase frequency dictionary (hereafter called cluster) yield those subthemes closely related to the main cluster theme.

Then, threshold values are assigned to the numerical indices. These indices are used to filter out the most closely related phrases to the cluster theme (e.g., see the example (TABLE 4-BOUNDARY LAYER-ABSTRACT DATABASE) following this section for part of a typical filtered cluster from the study).

TABLE 4

THEME PHRASE "BOUNDARY LAYER" - ABSTRACT DATABASE - SORT BY Ii

Cij	Ci	Ii (Cij/Ci)	Ij (Cij/Cj)	Eij (Ii*Ij)	CLUSTER MEMBER
14	18	0.778	0.022	0.0174	EXPANSION CORNER
7	9	0.778	0.011	0.0087	FULLY TURBULENT
6	8	0.750	0.010	0.0072	SEPARATION POINT
12	16	0.750	0.019	0.0144	LINEAR STABILITY THEORY
5	7	0.714	0.008	0.0057	BLEED CONFIGURATIONS

CODE:
Cij IS CO-OCCURRENCE FREQUENCY, OR NUMBER OF TIMES CLUSTER MEMBER APPEARS WITHIN +-50 WORDS OF CLUSTER THEME IN TOTAL TEXT;
Ci IS ABSOLUTE OCCURRENCE FREQUENCY OF CLUSTER MEMBER;
Cj IS ABSOLUTE OCCURRENCE FREQUENCY OF CLUSTER THEME;
Ii, THE CLUSTER MEMBER INCLUSION INDEX, IS RATIO OF Cij TO Ci;
Ij, THE CLUSTER THEME INCLUSION INDEX, IS RATIO OF Cij TO Cj,
AND Eij, THE EQUIVALENCE INDEX, IS PRODUCT OF INCLUSION INDEX BASED ON CLUSTER MEMBER Ii (Cij/Ci) AND INCLUSION INDEX BASED ON CLUSTER THEME Ij (Cij/Cj). Eij BEARS SOME SIMILARITY TO THE MUTUAL INFORMATION METHOD FROM COMPUTATIONAL LINGUISTICS, THAT COMPARES THE PROBABILITY OF TWO WORDS OCCURRING TOGETHER WITH THE PROBABILITY OF THE WORDS OCCURRING SEPARATELY.

For purposes of analysis, the cluster members in a given theme are segregated by their values of Inclusion Indices Ii and Ij. Ii is the ratio of Cij to Ci, and is the Inclusion Index based on the cluster member. Ij is the ratio of Cij to Cj, and is the Inclusion Index based on the theme word. Ii and Ij are categorized as either high or low. The dividing points between high and low Ii and Ij are the middle of the "knee" of the distribution functions of numbers of cluster members vs. values of Ii and Ij. All cluster members with Ii greater than or equal to approximately 0.5 were defined as having high Ii. All cluster members with Ij greater than or equal to 0.1 were defined as having high Ij. Table 5 presents examples of members in each of the Ii x Ij cells.

TABLE 5 - HIGH - LOW INCLUSION INDEX CELL MEMBERS

	Ij	
	LO	HI
SECONDARY INSTABILITY	!	CORNER

(Ci Ii Ij)	!	(Ci Ii Ij)
(4 1.000 .006)	!	(96 .583 .089)*
	!	
WALL HEAT FLUX	!	FLAT PLATE
HI (5 1.000 .008)	!	(115 .470 .086)*
	!	
BACKWARD-STEP INJECTOR	!	BLEED
(3 1.000 .005)	!	(97 .495 .077)*
Ii ----------		
HEAT TRANSFER	!	SHOCK WAVE
(218 .183 .064)	!	(275 .251 .110)
	!	
REYNOLDS NUMBER	!	LAYER
LO (132 .295 .062)	!	(1124 .243 .435)
	!	
TURBULENT FLOW	!	SHOCK
(57 .263 .024)	!	(1241 .197 .389)

CODE: THE FIRST NUMBER IN THE PARENTHESIS UNDER EACH PHRASE IS Ci (ABSOLUTE OCCURRENCE FREQUENCY IN TOTAL TEXT) FOR THE PHRASE, THE SECOND NUMBER UNDER EACH PHRASE IS Ii FOR THE PHRASE, AND THE THIRD NUMBER IN THE PARENTHESIS IS Ij. THE ASTERISK (*) FOLLOWING EACH PARENTHESIS IN THE HI Ii - HI Ij CELL DENOTES THAT EACH CELL MEMBER IS SLIGHTLY BELOW THE THRESHOLD.

A high value of Ii means that, whenever the cluster member appears in the total database text, there is a high probability that the theme phrase will appear within +-50 words of the cluster member. A high value of Ij means that, whenever the theme phrase appears in the total database text, there is a high probability that the cluster member will appear within +-50 words of the theme phrase.

Thus, phrases categorized as high Ii high Ij are coupled very strongly to the theme phrase. Whenever the theme phrase appears in the total database, there is a high probability that the cluster member will be physically close. Whenever the cluster member appears in the total database, there is a high probability that the theme phrase will be physically close. Whenever either phrase appears in the total database text, the other will be physically close.

There were no phrases in the present database that strictly met the threshold criteria for the high Ii high Ij cell. Past experience has shown that this cell tends to be the least populated, since those phrases which have a high Ii value tend to have a low Ij value, and vice versa. Some databases will have a few members in this cell, while others, such as the present database, have none.

The phrase FLAT PLATE in the high Ii high Ij cell of Table 5 reflects the reality that flat plates have been studied extensively in hypersonic/ supersonic flow because of their simple geometry, and one of the main focus areas on flat plate geometries has been boundary layer studies. Rationales for the presence of CORNER and BLEED are similar.

Consider phrases categorized as low Ii high Ij. Whenever the cluster member appears in the total database text, there is a low probability that it will be physically close to the theme phrase. Whenever the theme phrase appears in the total database text, there is a high probability that it will be physically close to the cluster member. This type of situation occurs when the frequency of occurrence of the cluster member Ci is substantially larger than the frequency of occurrence of the theme phrase Cj, and the cluster member and the theme phrase have some related meaning. Single word phrases have absolute frequencies an order of magnitude higher than double word phrases. Thus, the phrases categorized as low Ii high Ij are typically high frequency single word phrases. They are related to the theme phrase but much broader in meaning than the theme phrase. A small fraction of the time that these broad single word phrases appear, the more narrowly defined double word phrase theme will appear physically close. However, whenever the narrowly defined double word phrase theme appears, the broader related single word phrase cluster member will appear. The phrases under this heading can also be viewed as a higher level taxonomy of technical disciplines related to the theme.

The low Ii high Ij cell in Table 5 illustrates these concepts. Two of the three phrases are single words, and their absolute occurrence frequencies are relatively high. However, in the computer output, the members of this cell were ranked in decreasing order of Ij, and only two of the first forty phrases were non- single words. Also, as stated in the previous paragraph, the more narrowly defined double word phrase SHOCK WAVE was preceded by the broader related single word phrase SHOCK.

Consider phrases categorized as high Ii low Ij. Whenever the cluster member appears in the total database text, there is a high probability that it will be physically close to the theme phrase. Whenever the theme phrase appears in the total database text, there is a low probability that it will be physically close to the cluster member. This type of situation occurs when the frequency of occurrence of the cluster member Ci is substantially smaller than the frequency of occurrence of the theme phrase Cj, and the cluster member and the theme phrase have some related meaning. Thus, the phrases categorized as high Ii low Ij tend to be low frequency double and triple word phrases, related to the theme phrase but very narrowly defined.

A large fraction of the time that these very narrow double and triple word phrases appear, the relatively broader double word phrase theme will appear physically close. However, a small fraction of the time that the relatively broad double word phrase theme appears, the more narrow double and triple word phrase cluster member will appear. This grouping has the potential for identifying "needle-in-a-haystack" type thrusts which occur infrequently but strongly support the theme when they do occur. For studies whose main focus is innovation and discovery through related literatures [Kostoff, 1999], this grouping would be central in linking strongly related literatures. One of many advantages of full text over key or index phrases is this illustrated ability to retain low frequency but highly important phrases, since the key phrase approach ignores the low frequency phrases.

The phrases in the high Ii low Ij cell in Table 5 are multi- words, and contain substantially more detailed technical content than those in the low Ii cells. All the high Ii low Ij phrases had very low absolute occurrence frequencies, and this order of magnitude is typical of the high Ii phrases (especially Ii of unity) in all the databases examined so far.

Finally, consider phrases categorized as low Ii low Ij. Whenever the cluster member appears in the total database text, there is a low probability that it will be physically close to the theme phrase. Whenever the theme phrase appears in the total database text, there is a low probability that it will be physically close to the cluster member. The phrases in this category are intermediate to the high technical content low frequency high Ii low Ij typically triple word phrases and the low technical content high frequency low Ii high Ij typically single word phrases. Thus, the low Ii low Ij phrases tend to have moderate technical content, moderate frequency, and typically double word structure. These phrases tend to be most useful in providing a balanced technical description of the relationships within the cluster of interest; they are not too detailed to obscure the larger context, yet they are not too general to be devoid of meaningful content. For example, using only the three phrases in low Ii low Ij cell in Table 5 to describe the message from the proximity analysis for boundary layers in hypersonic flows, much of the research effort appears to be focused on HEAT TRANSFER in TURBULENT FLOWS at high REYNOLDS NUMBER. While the field of hypersonic research is far more detailed and complex than this brief statement implies, a more coherent and meaningful summary description is possible with low Ii low Ij phrases than with the overly detailed high Ii low Ij phrases or the overly generic low Ii high Ij phrases. The phrases in each of the four cells shown have their unique roles to play in the description and analysis of the proximity relationships, and the preceding sections have shown the proper matching of cells to analyses desired.

4.5.2 Analysis

The full text database was split into two databases. One was the abstract narrative database (referred to as ABSTRACT in the phrase proximity analysis below), and phrase proximity analysis of this database yielded mainly topical theme relationships. The other database (referred to as BLOCK below) consisted of records (one for each published paper) containing four fields: author(s), title, journal name, author(s) institutional address(es). Phrase proximity analysis of this database yielded not only topical theme relationships from the proximal title words, but also relationships between technical themes and authors, journals, and institutions.

Because of space limitations in this document, only one theme, BOUNDARY LAYER, was chosen for the phrase proximity analysis. It was high frequency in both the abstracts and titles, and is a central theme in hypersonic flow over aerodynamic bodies. In the following section, the cluster theme BOUNDARY LAYER is analyzed for the BLOCK and ABSTRACT database components. Further, for each of these database components, the cluster theme is analyzed from the two perspectives of high Ii low Ij and low Ii high Ij. The phrase proximity analysis process for BOUNDARY LAYER consisted of providing the expert with two lists of cluster members, one sorted by Ii and the other by Ij. By visual examination of these lists, the expert constructed categories of related items, and these relationships are reported below.

The types and numbers of categories possible are limited by the perceptual capabilities of the expert(s), and could vary substantially among experts. This issue of category definition is a good example of the advantages and challenges of the full text procedure reported in this paper relative to the key word or index word approaches used in most co-word based analyses. Full text provides many more degrees of freedom relative to index words, and therefore many more possibilities of different relational categories. However, analysts with the ability to perceive large numbers of

relationships, especially the highest value relationships, are required to obtain maximal benefit from the increased degrees of freedom.

Possibly, in the future, a 'tagging' could be applied initially to each cluster member to place it in a number of potential categories before the actual categorization procedure starts. Then, a computerized selection process would generate the category placement. This approach would still require the perceptual capabilities of the analyst/ expert to perform the initial 'tagging' expansively, but might prove more systematic and comprehensive in providing a broader, more complete, range of categories. It would, however, be very labor intensive, especially for the larger clusters.

4.6 Phrase Proximity Analysis - BOUNDARY LAYER

4.6.1 BLOCK database; low Ii high Ij.
The phrases describe the more generic associations with BOUNDARY LAYER.

Major Associated Journals and Conferences

AIAA JOURNAL, JOURNAL OF FLUID MECHANICS, AMERICAN SOCIETY OF MECHANICAL ENGINEERS, ISVESTIYA AN SSSR, MEKHANIKA ZHIDKOSTI I GAZA

Major Associated Institutions

ASME, AEROSPATIALE, UNIV OF MANCHESTER, NII MEKHANIKI MGU

Major Associated Countries

USA, JAPAN, RUSSIA, FRANCE, ENGLAND, ITALY

Broad Associated Technical Themes

FLUID FLOW, AERODYNAMICS, HYDRODYNAMICS, NUMERICAL METHODS, APPLIED PHYSICS, HEAT TRANSFER, SHOCK WAVES, AIRCRAFT, MATHEMATICAL MODELS, THERMODYNAMICS, TURBULENCE, TEMPERATURE, WIND TUNNELS, TRANSITION, MACH NUMBER, LAMINAR FLOW, VISCOUS FLOW, GAS DYNAMICS

4.6.2 BLOCK database; high Ii low Ij.
The phrases describe the more specific associations with BOUNDARY LAYER.

Major Associated Authors

AUDIFFREN-N, ARNETTE-SA, SMITS-AJ, TAKAAKI-S, DUDIN-GN, SHINJI-H, KIMIO-S, HAMED-A, FUMIO-H, HORSTMAN-CC

Major Associated Institutions

UNIV OF MANCHESTER, U-BARI, SANDIA NATIONAL LAB, PRINCETON UNIV,

DLR INSTITUTE, UNIV OF TEXAS-ARLINGTON, OLD DOMINION UNIV, NII MEKHANIKI MGU, ECOLE POLYTECHNIQUE FEDERAL DE LAUSANNE

Major Associated Solid Body Topographies

SHARP FINS, CONCAVE SURFACE CURVATURE, SWEPT WING, FLAT PLATE

Major Associated Flow Field Phenomena

WAVES (SWEPT SHOCK WAVE, OBLIQUE INSTABILITY WAVE, CROSSING SHOCK, NORMAL SHOCK WAVE, FIRST-MODE WAVES), INVISCID GORTLER VORTICES, ADVERSE PRESSURE GRADIENTS, REYNOLDS STRESSES, ISOTROPIC TURBULENCE, VISCOUS FLOW, COMPRESSIBLE BOUNDARY LAYER, LAMINAR FLOW, FLOW SEPARATION, TURBULENT COMPRESSIBLE FLOW, BOUNDARY LAYER TRANSITION

There is an absence of terms relating to the real gas effects which appear at very high speeds, such as dissociation, ionization, and radiation, which leads to the interpretation that very little is contained in this database relative to very high Mach number flows. A sampling of the detailed database records confirms this observation.

Major Associated Variables

ELECTRON DENSITY, HEAT TRANSFER COEFFICIENTS, REYNOLDS NUMBER, MACH NUMBER, PRANDTL NUMBER

Major Computational Methods for Determining these Variables

COUPLED EULER, SPECTRAL METHOD, COMPUTER SIMULATION, LINEAR STABILITY THEORY, ASYMPTOTIC THEORY, ALGEBRAIC TURBULENCE MODEL, BOUNDARY LAYER EQUATIONS, NAVIER-STOKES EQUATIONS

The only experimental facility or method mentioned is WIND TUNNELS. The relative absence in this section of experimental methods for determining these variables reflects the growing trend toward replacing expensive fluid dynamics experiments, especially high speed experiments which may require large and expensive facilities or expensive flight experiments, with less expensive computational fluid dynamics techniques.

4.6.3 ABSTRACT database; low Ii high Ij.
These phrases describe the more generic associations with BOUNDARY LAYER.

Major Flow Regimes

TURBULENT, VISCOUS, INVISCID, LAMINAR, COMPRESSIBLE

Major Flowfield Phenomena

SHOCK WAVE, SEPARATION, TRANSITION, HEAT, EXPANSION, STABILITY, INSTABILITY, NONEQUILIBRIUM, SHEAR, VORTICES, EQUILIBRIUM, REATTACHMENT, ELECTRON, KINETICS, OSCILLATION, FRICTION

Major Flowfield Variables

PRESSURE, TEMPERATURE, VELOCITY, STRESS, HEAT TRANSFER, and the dimensionless variables of MACH NUMBER, REYNOLDS NUMBER

Major Topographies

FLAT PLATE, AXISYMMETRIC, BLUNT BODY, TWO-DIMENSIONAL, THREE-DIMENSIONAL, SHARP

Major Regions

SURFACE, WALL, DOWNSTREAM, CORNER, NOZZLE, JET, UPSTREAM, EDGE, STAGNATION

4.6.4 ABSTRACT database; high Ii low Ij.

These phrases describe the more generic associations with BOUNDARY LAYER.

Major Flowfield Phenomena

shock (SWEPT SHOCK WAVE, EXTERNALLY GENERATED SHOCK, SHOCK MOTION, POST-SHOCK EXPANSION, SHOCK WAVE TURBULENT, SHOCK GENERATOR),

separation (STRONG ADVERSE PRESSURE GRADIENT, SEPARATION SHOCK OSCILLATION, TRANSITION ONSET, TRANSITION DETECTION, SEPARATION BUBBLES, INFLECTION POINTS, SEPARATION POINT, REATTACHMENT),

viscous/ inviscid (INVISCID MODE, INVISCID BREAKDOWN, INVISCID BOUNDARY LAYER, VISCOUS-INVISCID INTERACTION, VISCOUS SUBLAYER, VISCOUS BOUNDARY LAYER),

stability/ instability (SECONDARY INSTABILITY MECHANISM, LINEAR STABILITY THEORY, WALL PRESSURE FLUCTUATIONS, ACOUSTIC MODES, ACOUSTIC RESPONSES),

chemistry (NITRIC OXIDE PRODUCTION, FINITE SURFACE CATALYSIS, VIBRATIONAL KINETICS),

compression/ expansion (COMPRESSIBLE BOUNDARY LAYERS, CENTERED EXPANSION, BULK COMPRESSION, TWO-DIMENSIONAL COMPRESSION, TURNING

ANGLE),

turbulent/ laminar (ROUGH WALL ALGEBRAIC TURBULENCE, TAYLOR-GORTLER VORTICES, ELONGATED LONGITUDINAL STRUCTURES, IRREVERSIBILITIES, FULLY LAMINAR, FULLY TURBULENT, MIXING LENGTH, LARGE SCALE STRUCTURES, LENGTH SCALES),

integrated effects (BOUNDARY LAYER PROFILES, BOUNDARY LAYER DISPLACEMENT, BOUNDARY LAYER THICKNESS),

other (KINETIC LOSSES, HARMONIC POINT SOURCE, REYNOLDS STRESSES)

Major Measured Variables

WALL HEAT FLUX, FREESTREAM REYNOLDS NUMBER, FLUCTUATING PRESSURE FIELD, CONCAVE STREAMLINE CURVATURE, BLEED RATES, STREAMLINE REYNOLDS NUMBER, MOMENTUM THICKNESS, TOTAL ENTHALPY, STRESS TENSOR, ELECTRON CONCENTRATION, HEAT TRANSFER COEFFICIENTS, MEAN VELOCITY

Major Topographies

SLENDER CONES, FLAT PLATE, FORWARD-FACING RAMP, SHARP FINS

Major Body Regions

EXPANSION SURFACES, CONVEX CORNER, COMPRESSION CORNERS, AXISYMMETRIC QUIET NOZZLE, CORNER ANGLES, EXPANSION CORNER

Major Boundary Modifications

MASS REMOVAL BLEED SLOT, BACKWARD-STEP INJECTOR, AIR INTAKE, SUCTION

Major External Boundary Conditions

ROUGH WALL, HIGH ALTITUDE HYPERSONIC FLIGHT, WALL CURVATURE, FLEXIBLE SURFACE, LARGE REYNOLDS NUMBER, FREE STREAM MACH NUMBER

Major Computational and Experimental Approaches and Models

PRESSURE LAW, SOLVING THE LINEARIZED, MODEL OF CEBICI, COUPLES EULER, SURFACE OIL FLOW VISUALIZATIONS, REMESHING, PRESSURE SURVEY, NAVIER STOKES SOLVERS, FULL NAVIER STOKES CODE, LINEARIZED EQUATIONS, LINEART STABILITY THEORY, K-EPSILON MODEL, PITOT PRESSURE, TRIPLE DECK

Major Applications

ATMOSPHERIC REENTRY, HYPERSONIC FLIGHT REGIME, PEGASUS.

Due to the limited size of the hypersonic database, some of the entries in the BLOCK database analysis are somewhat sparse. As an example of the breadth of results available from a larger database, see the output from the near-earth space study presented for the key phrase REMOTE SENSING [Kostoff, 1998a].

4.7 Taxonomies

The different types of DT outputs allow different types of taxonomies, or classifications into component categories, to be generated. Such categorizations, anologous to the independent axes of a mathematical coordinate system, allow the underlying structure of a field to be portrayed more clearly, leading to more focused analytical and management analyses. Two separate taxonomies will be discussed here, but due to space limitations, only one will be presented.

The first taxonomy derives from the phrase frequencies. The authors examined the phrase frequency outputs, then arbitrarily grouped the high frequency phrases into different, relatively independent, categories for which all remaining terms would be accounted. One taxonomy was developed for the SCI phrase frequencies, and another taxonomy for the EC phrase frequencies. The second taxonomy derives from the phrase frequency and proximity analyses, and will be presented here. From the phrase frequency analysis, about sixty high frequency technical phrases were identified by the domain expert as pervasive themes. A proximity analysis was done for each of these high frequency phrases. A phrase frequency dictionary, or cluster, was generated for each phrase. This cluster contained those phrases which were in close physical proximity to the pervasive theme throughout the text. The cluster generation process described in this paragraph, and many of the clusters obtained, were also used for the analysis of the theme relationships presented in section 4.5 (Phrase Proximity Analysis -Relationships Among Themes)

The degree of overlap among clusters was computed by counting the number of shared phrases, in the following manner. The cluster sizes were normalized, and each normalized cluster pair which shared more than a threshold number of common phrases was viewed as overlapping. These overlapping clusters were viewed as links in a chain, with the different chains being relatively independent. Each chain was then defined as a megacluster, or category of the larger taxonomy. The threshold level for cluster overlap was determined by varying the number of common phrases parametrically, and observing the patterns of aggregation of the clusters. The parametric variation of these patterns of aggregation is shown in Table 6. The leftmost column is the cluster, or theme, name. The numeric column headings (e.g., 60, 50) represent the number of overlaps, or common phrases, among normalized clusters. The alphanumeric matrix entries are the members of the different chains. The alphabetical characters of the matrix entry identify the chain, and the numerical characters of the matrix entry identify the minimum number of overlaps.

TABLE 6
FORMATION OF MEGATHEMES

THEME NAME / #OVERLAPS	60	50	40	35	30	25	20	15

SHOCK TUNNEL	A60	A50	A40	A35	ABC30	ABCDE25	ABCDEFGH20	ABCDEFGHIJK15
HIGH TEMPERATURE	A60	A50	A40	A35	ABC30	ABCDE25	ABCDEFGH20	ABCDEFGHIJK15
HYPERSONIC FLIGHT		A50	A40	A35	ABC30	ABCDE25	ABCDEFGH20	ABCDEFGHIJK15
BOUNDARY LAYERS			B40	B35	ABC30	ABCDE25	ABCDEFGH20	ABCDEFGHIJK15
OBLIQUE SHOCK			B40	B35	ABC30	ABCDE25	ABCDEFGH20	ABCDEFGHIJK15
REYNOLDS NUMBER			B40	B35	ABC30	ABCDE25	ABCDEFGH20	ABCDEFGHIJK15
TRAILING EDGE			B40	B35	ABC30	ABCDE25	ABCDEFGH20	ABCDEFGHIJK15
SUPERSONIC AIRCRAFT				B35	ABC30	ABCDE25	ABCDEFGH20	ABCDEFGHIJK15
PRESSURE GRADIENT				B35	ABC30	ABCDE25	ABCDEFGH20	ABCDEFGHIJK15
SUPERSONIC SPEEDS				B35	ABC30	ABCDE25	ABCDEFGH20	ABCDEFGHIJK15
TURBULENT BOUNDARY LAYER				B35	ABC30	ABCDE25	ABCDEFGH20	ABCDEFGHIJK15
VISCOUS SHOCK LAYER				C35	ABC30	ABCDE25	ABCDEFGH20	ABCDEFGHIJK15
PERFECT GAS				C35	ABC30	ABCDE25	ABCDEFGH20	ABCDEFGHIJK15
SURFACE PRESSURE					ABC30	ABCDE25	ABCDEFGH20	ABCDEFGHIJK15
SUPERSONIC NOZZLE EXIT				D35	D30	ABCDE25	ABCDEFGH20	ABCDEFGHIJK15
NUMERICAL SIMULATION				D36	D30	ABCDE25	ABCDEFGH20	ABCDEFGHIJK15
WIND TUNNELS				D37	D30	ABCDE25	ABCDEFGH20	ABCDEFGHIJK15
FLOW CONDITIONS				D38	D30	ABCDE25	ABCDEFGH20	ABCDEFGHIJK15
GAS FLOW				D39	D30	ABCDE25	ABCDEFGH20	ABCDEFGHIJK15
HYPERSONIC VEHICLES					D30	ABCDE25	ABCDEFGH20	ABCDEFGHIJK15
SUPERSONIC COMBUSTION					D30	ABCDE25	ABCDEFGH20	ABCDEFGHIJK15
GROWTH RATE				E35	E30	ABCDE25	ABCDEFGH20	ABCDEFGHIJK15
SHEAR LAYERS				E35	E30	ABCDE25	ABCDEFGH20	ABCDEFGHIJK15
SUPERSONIC JETS				E35	E30	ABCDE25	ABCDEFGH20	ABCDEFGHIJK15
CONVECTIVE MACH					E30	ABCDE25	ABCDEFGH20	ABCDEFGHIJK15
LARGE SCALE						ABCDE25	ABCDEFGH20	ABCDEFGHIJK15
SUPERSONIC FLOWS						ABCDE25	ABCDEFGH20	ABCDEFGHIJK15
DIRECT SIMULATION		F50	F40	F35	F30	F25	ABCDEFGH20	ABCDEFGHIJK15
MONTE CARLO		F50	F40	F35	F30	F25	ABCDEFGH20	ABCDEFGHIJK15
FLOW VISUALIZATION						G25	ABCDEFGH20	ABCDEFGHIJK15
PRESSURE MEASUREMENTS						G25	ABCDEFGH20	ABCDEFGHIJK15
ANGLE OF ATTACK						H25	ABCDEFGH20	ABCDEFGHIJK15
BLUNT BODY						H25	ABCDEFGH20	ABCDEFGHIJK15
BOW SHOCK							ABCDEFGH20	ABCDEFGHIJK15
MIXING LAYER			I40	I35	I30	I25	I20	ABCDEFGHIJK15
SUPERSONIC MIXING			I40	I35	I30	I25	I20	ABCDEFGHIJK15
CHEMICAL REACTION			J40	J35	J30	J25	JK20	ABCDEFGHIJK15
HEAT RELEASE			J40	J35	J30	J25	JK20	ABCDEFGHIJK15
MASS FLOW				K35	K30	K25	JK20	ABCDEFGHIJK15
NUMERICAL SOLUTIONS				K35	K30	K25	JK20	ABCDEFGHIJK15
COMPUTATIONAL FLUID DYNAMICS								ABCDEFGHIJK15
FREESTREAM MACH								ABCDEFGHIJK15
PRESSURE DISTRIBUTION								ABCDEFGHIJK15
STATIC PRESSURE								ABCDEFGHIJK15
TOTAL PRESSURE								ABCDEFGHIJK15
EXPERIMENTAL STUDY							L20	L15
HEAT FLUX							L20	L15
EULER EQUATIONS							M20	M15
FINITE VOLUME							M20	M15
BOUNDARY CONDITIONS								
FINITE ELEMENT								

For example, in the first column (60), there is one chain (A). It has two links/ themes/ clusters (SHOCK TUNNEL, HIGH TEMPERATURE), and the themes/ clusters (normalized to 100 cluster component phrase members) have at least 60 phrase members in common. As another example, in the column with heading 30, there are seven chains (ABC, D, E, F, I, J, K), and every link/ theme/ cluster in each chain has at least thirty phrase members in common with at least one other link in the chain. The largest chain (ABC) is an amalgamation of three component chains ((A, B, C) which were formed previously. One value of following the chain formations parametrically is that the strong link associations evidenced by the component chains A, B, C can be readily identified. Obviously, many taxonomies are possible with this approach, depending on the final threshold value selected. If the threshold value is set too high (e.g., >60), there will be a large number of independent categories, and the taxonomy will be unwieldy for any practical use. If the threshold value is set too low (e.g., <15), all the categories tend to merge, and the taxonomy does not provide much information. The results from Table 6, modified by the judgement and experience of the authors, are used to form the useful taxonomy shown in Table 7. The phrases preceded by an asterisk (*) are the megacluster themes, and the phrases preceded by a hyphen (-) are their component cluster themes.

This taxonomy reflects very accurately the thrust areas of hypersonic and supersonic flow over aerodynamic bodies. The HYPERSONIC EXPERIMENTS measurements emphasize high pressure and temperature conditions, focusing on heat flux data and flow visualization techniques, and making increased use of shock tunnels relative to free flight experiments. The COMPUTATIONAL FLUID DYNAMICS approaches, which are assuming a greater portion of HSF research, encompass finite volume and finite element techniques and monte carlo simulations as well. The three post-shock regions of SHOCK LAYER, SHEAR AND MIXING LAYER, and BOUNDARY LAYER, each constitute emphasis areas with unique sub-thrust areas. As the threshold conditions for overlapping phrases were further reduced from the values reflective of Table 6, these three areas shortly merged into one, paralleling their intrinsic physical connectivity. TURBULENT FLOW with its high mixing and heat flux rates is of primary interest, while ASYMMETRICAL FLOW with its potentially higher lift coefficients assumes increasing importance for improving hypersonic vehicle performance. NOZZLE FLOW has a dual importance: the study and control of high speed flow from actual aircraft and missile nozzle exits to maximize thrust and minimize fuel consumption, and similar studies of laboratory nozzle flows to understand the flowfield fluid dynamics and improve the nozzle as a high speed flow source. Finally, INTERNAL ENERGY PRODUCTION is important for studying high speed combustion, as well as the reaction and dissociation chemistry of high speed gases.

TABLE 7

HSF TAXONOMY - MEGACLUSTERS

*HYPERSONIC EXPERIMENTS *COMPUTATIONAL FLUID DYNAMICS

-SHOCK TUNNEL
DYNAMICS
-HYPERSONIC FLIGHT
-HIGH TEMPERATURE
-FLOW VISUALIZATION
-PRESSURE MEASUREMENTS
-STATIC PRESSURE
-EXPERIMENTAL STUDY
-HEAT FLUX

*BOUNDARY LAYER
-BOUNDARY LAYERS
-SURFACE PRESSURE
-PRESSURE GRADIENT
-OBLIQUE SHOCK
-REYNOLDS NUMBER
-TRAILING EDGE
-SUPERSONIC FLOWS

*SHEAR AND MIXING LAYER
-SHEAR LAYERS
-TOTAL PRESSURE
-SUPERSONIC JETS
-GROWTH RATE
-CONVECTIVE MACH
-FREESTREAM MACH
-MIXING LAYER
-SUPERSONIC MIXING

*TURBULENT FLOW
-TURBULENT FLOW
-TURBULENT BOUNDARY LAYER
-TURBULENCE MODEL
-LARGE SCALE

*INTERNAL ENERGY PRODUCTION
-HEAT RELEASE
-CHEMICAL REACTION
-MASS FLOW
-NUMERICAL SOLUTIONS

-COMPUTATIONAL FLUID
DYNAMICS
-DYNAMIC PRESSURE
-DIRECT SIMULATION
-MONTE CARLO
-EULER EQUATIONS
-FINITE VOLUME
-BOUNDARY CONDITIONS
-FINITE ELEMENT

*SHOCK LAYER
-SHOCK LAYER
-VISCOUS SHOCK LAYER
-PERFECT GAS
-SUBSONIC AND SUPERSONIC
-SUPERSONIC AIRCRAFT
-SUPERSONIC SPEEDS

*NOZZLE FLOW
-NOZZLE EXIT
-SUPERSONIC NOZZLE
-FLOW CONDITIONS
-GAS FLOW
-PRESSURE DISTRIBUTION
-NUMERICAL SIMULATION
-HYPERSONIC VEHICLES
-SUPERSONIC COMBUSTION
-WIND TUNNELS

*ASYMMETRICAL FLOW
-ANGLE OF ATTACK
-BLUNT BODY
-BOW SHOCK

5. CONCLUSIONS

Combination of the average of 2.63 authors per HSF paper and the SCI keywords' emphasis on computational approaches implies that many of the HSF research projects are individual author or small team CFD efforts. There was a concentration of output in the top authors, journals,

institutions, and countries. The top authors had an order of magnitude larger number of listings than the average, as did the top journals and top institutions.

The top handful of SCI countries accounted for most of total addresses, and the dominance of these few countries seemed to be pervasive throughout the different research areas. Most of the cited authors were cited once, and far fewer were cited twice. A relatively few percent received large numbers of citations. However, the most cited authors, while prolific, are not the most prolific authors, and vice versa. Again, most of the cited papers received only one citation and far less received two citations. Relatively few papers were highly cited. From the citation year results, the most recent papers are the most highly cited. This reflects a rapidly evolving field of research. Most sources were cited only once and almost an order of magnitude less were cited twice. Relatively few sources were highly cited. There is a moderately stronger correlation between journals which are highly cited and which contain large numbers of HSF papers than between highly prolific and cited authors.

Publication and citation frequency distribution results were presented for authors, papers, journals, and organizations for six different topical fields (Space, Chemistry, HSF, Hydrodynamics, Aircraft, Fullerenes) from six different studies. Most of the distributions transcended the different topical fields, appearing to be topic-independent, and differed modestly for the type of distribution function (author, journal, etc.).

The SCI keywords tend to emphasize the more fundamental aspects of hypersonic/ supersonic flow (INSTABILITY WAVES, NAVIER- STOKES EQUATIONS, HYPERBOLIC CONSERVATION LAWS, NUMERICAL SIMULATIONS, ASYMPTOTIC ANALYSIS, INDUCED IODINE FLUORESCENCE, MACH REFLECTION), while the EC keywords tend to emphasize the more applied aspects of hypersonic/ supersonic flow (WIND TUNNELS, SUPERSONIC AIRCRAFT, HEAT TRANSFER, AERODYNAMIC LOADS, PRESSURE EFFECTS, RAMJET ENGINES, LIFT, SPACECRAFT). The SCI keywords also emphasize computational approaches, thereby substantiating the conclusion reached above that hypersonics research has a large individual investigator or small team component focusing on CFD. The EC keywords are more inclusive of experimental methods.

From a global perspective, the SCI database portrays the major focus of HSF to be SUPERSONIC/ HYPERSONIC MACH NUMBER flows over simple shapes (FLAT PLATE LEADING EDGE) at ANGLES OF ATTACK containing BOW SHOCKS and OBLIQUE SHOCKS. The experimental focus is WIND TUNNEL TESTS with measurement emphasis on SURFACE HEAT TRANSFER and SURFACE PRESSURE DISTRIBUTIONS within the VISCOUS SHOCK LAYER, SHEAR LAYER, SUPERSONIC MIXING LAYER, and TURBULENT BOUNDARY LAYER; the analytical focus is NUMERICAL SIMULATION by COMPUTATIONAL FLUID DYNAMICS with FINITE ELEMENT and MONTE CARLO SIMULATION, using the COMPRESSIBLE NAVIER-STOKES EQUATIONS to model the VISCOUS SHOCK LAYER and near-body region, and using the EULER EQUATIONS to model the outer inviscid region. Conspicuous by their absence are exotic gas mixtures (helium, hydrogen, etc) which would simulate other planetary atmospheres, exotic body shapes which would simulate novel vehicle designs, and real gas effects (dissociation, ionization, radiation, etc) which would accompany very high Mach numbers characteristic of planetary entry speeds.

This paper has presented a number of advantages of using DT and bibliometrics for deriving technical intelligence from the published literature. Large amounts of data can be accessed and analyzed, well beyond what a finite group of expert panels could analyze in a reasonable time period. Preconceived biases tend to be minimized in generating roadmaps. Compared to standard co-word analysis, DT uses full text, not index words, and can make maximum use of the rich semantic relationships among the words. It also has the potential of identifying low occurrence frequency but highly theme related phrases which are 'needles-in-a-haystack', a capability unavailable to any of the other co-occurrence methods. Combined with bibliometric analyses, DT identifies not only the technical themes and their relationships, but relationships among technical themes and authors, journals, institutions, and countries. Unlike other roadmap development processes, DT generates the roadmap in a 'bottom-up' approach. Unlike other taxonomy development processes, DT can generate many different types of taxonomies (because it uses full text, not key words) in a 'bottom-up' process, not the typical arbitrary 'top-down' taxonomy specification process. Compared to co-citation analysis, DT can use any type of text, not only published literature, and it is a more direct approach to identifying themes and their relationships. The maximum potential of the DT and bibliometrics combination is achieved when these two approaches are combined with expert analysis of selected portions of the database. If a manager, for example, wants to identify high quality research thrusts as well as science and technology gaps in specific technical areas, then an initial DT and bibliometrics analysis will provide a contextual view of work in the larger technical area; i.e., a strategic roadmap. With this strategic map in hand, the manager can then commission detailed analysis of selected abstracts to assess the quality of work done as well as identify work which needs to be done (promising opportunities).

6. REFERENCES

Anwar, M. A., and Abubakar, A. B. (1997). Current State of Science and Technology in the Muslim World. Scientometrics. 40:1, pp 23-44

ARPA. (1994). Wingship Investigation. Volume 3. Technology Roadmap. Final report. Advanced Research Projects Agency, Arlington, VA. 208p.

Arsham, H. (1993). Managing Project Activity Duration Uncertainties. Omega-International Journal Of Management Science. Vol 21, Iss 1, pp 111-122.

Attar, R. and Fraenkel, A.S. (1977). Local Feedback in Full- Text Retrieval Systems. Journal of the ACM. 24:3.

Barker, D. and Smith, D. (1995). Technology Foresight Using Roadmaps. Long Range Planning. 28:2. pp.21-29.

Bates, M. J. (1986) Subject Access in On-Line Catalogs: A Design Model. JASIS. 37:6.

Bauin, S. (1986). Aquaculture: A Field by Bureaucratic Fiat.in: Callon, M. Law, J. and Rip, A. (eds.) Mapping the Dynamics of Science and Technology. Macmillan Press Ltd. London.

Blau, J.R. (1978). Sociometric Structure of a Scientific Discipline. in: Jones, R.A. (ed.) Research in the Sociology of Knowledge, Sciences, and Art: An Annual Compilation of Research I, JAI Press, Greenwich, CT.

Braam, R. Moed, H. and Van Raan, A. (1991a),(1991b). Mapping of Science by Combined Co-Citation and Word Analysis. 1. Structural Aspects. Journal of the American Society for Information Science. 42 (4). Mapping of Science by Combined Co- Citation and Word Analysis. 2. Dynamical Aspects. Journal of the American Society for Information Science. 42 (4).

Bradford, S. C. (1934) Sources of Information on Specific Subjects. Engineering. 137.

Callan, J., Croft, W.B., and Broglio, J. (1995) TREC and TIPSTER Experiments with INQUERY. Information Processing and Management. 31:3.

Callon M., Courtial, J. P., and Turner, W. A. (1979) PROXAN: A Visual Display Technique for Scientific and Technical Problem Networks. Second Workshop on the Measurement of R&D Output. Paris, France. December 5-6.

Callon, M. Courtial, J.P. Turner, W.A. and Bauin, S. (1983). From Translations to Problematic Networks: An Introduction to Co- word Analysis. Social Science Information. 22.

Callon, M. (1986). Pinpointing Industrial Invention: An Exploration of Quantitative Methods for the Analysis of Patents. in: Callon, M.Law, J. and Rip, A. (eds.) Mapping the Dynamics of Science and Technology. Macmillan Press Ltd., London.

Callon, M., Courtial, J. P., and Laville, F. (1991). Co-Word Analysis As a Tool for Describing the Network of Interactions Between Basic and Technological Research - The Case of Polymer Chemistry. SCIENTOMETRICS. 22:1, pp 155-205.

Chen, H. C., et al. (1997) A Concept Space Approach to Addressing the Vocabulary Problem in Scientific Information Retrieval: An Experiment on the Worm Community System. JASIS. 48:1.

Chen, H. C., et al. (1998) Alleviating Search Uncertainty through Concept Associations: Automatic Indexing, Co-occurrence Analysis, and Parallel Computing. JASIS. 49:3.

Croft, W., and Harper, D. (1979) Using Probabilistic Models of Document Retrieval Without Relevance Information. Journal of Documentation. 35.

Crouch, C. J. (1990) An Approach to the Automatic Construction of Global Thesauri. Information Processing and Management. 26:5. 1990.

De Saussure, F. (1949). Cours de Linguistique Generale. 4eme Edition, Librairie Payot, Paris.

Dodin-B.M.; Elmaghraby-S.E. (1985). Approximating the Criticality Indices of the Activities in PERT Networks. North Carolina State Univ. at Raleigh, Operations Research, Report Number ARO1635218MA, 19p.

Doyle, L. B. (1962) Indexing and Abstracting by Association. American Documentation. 13:4.

Engelsman, E. C. and Van Raan, A. F. J. (1991). Mapping of Technology: A First Exploration of Knowledge Diffusion amongst Fields of Technology. Research Report to the Ministry of Economic Affairs, CWTS-91-02, Centre for Science and Technology Studies. Leiden.

Furnas, G. W., et al (1987) The Vocabulary Problem in Human- system Communication. Communications of the ACM. 30:11.

Georghiou, L. Giusti, W.L. Cameron, H.M. and Gibbons, M. (1988). The Use of Co-nomination Analysis in the Evaluation of Collaborative Research. in Van Raan, A.F.J. (ed.), Handbook of Quantitative Studies of Science and Technology. North Holland.

Gill, S. P., Frye, P. E., Littman, F. D., and Meisl, C. J. (1994). Power Systems for Future Missions. Final Report. Rockwell International Corp. Canoga Park, CA. Rocketdyne Div. Report Number NAS126195320, E8735, NASACR195320, 84p.

Gomez, L. M., Lochbaum, C. C., and Landauer, T. K. (1990) All the Right Words: Finding What You Want as a Function of the Indexing Vocabulary. JASIS. 41.

Groenveld, P. (1997). Roadmapping Integrates Business and Technology. Research-Technology Management. Sept-Oct.

Healey, P. Rothman, H. and Hoch, P. (1986). An Experiment in Science Mapping for Research Planning. Research Policy. 15.

Hornby, A.S. Gatenby, E. V. and Wakefield, H. (1942). Idiomatic and Syntactic English Dictionary. Kaitakusha, Tokyo, Japan.

Jing, Y. and Croft, W.B. (1994) An Association Thesaurus for Information Retrieval. Proceedings of RIAO 94.

Kostoff, R. N. (1991). Database Tomography: Multidisciplinary Research Thrusts from Co-Word Analysis. Proceedings: Portland International Conference on Management of Engineering and Technology. October 27-31.

Kostoff, R. N. (1992) Research Impact Assessment. Proceedings: Third International Conference on Management of Technology. Miami, FL, February 17-21. Larger text available from author.

Kostoff, R. N. (1993a). Database Tomography for Technical Intelligence. Competitive Intelligence Review. 4:1.

Kostoff, R. N. (1993b) Co-Word Analysis. in Assessing R&D Impacts: Method and Practice. Bozeman, B. and Melkers, J., Eds. (Kluwer Academic Publishers, Norwell, MA).

Kostoff, R. N. (1994a). Research Impact Quantification. R&D Management. 24:3.

Kostoff, R.N. (1994b). Database Tomography: Origins and Applications. Competitive Intelligence Review 5:1. Spring.

Kostoff, R. N., Eberhart, H. J., and Miles, D. (1995). System and Method for Database Tomography. U. S. Patent Number 5440481. August 8.

Kostoff, R. N. (1997a). Database Tomography for Information Retrieval. Journal of Information Science. 23:4.

Kostoff, R. N. (1997b). Database Tomography for Technical Intelligence: Comparative Roadmaps of the Research Impact Assessment Literature and the Journal of the American Chemical Society. Scientometrics. 40:1.

Kostoff, R. N. (1997c). Use and Misuse of Metrics in Research Evaluation. Science and Engineering Ethics. 3:2.

Kostoff, R. N. (1998a). Database Tomography for Technical Intelligence: A Roadmap of the Near-earth Space Science and Technology Literature. Information Processing and Management, 34:1.

Kostoff, R. N. (1998b). Science and Technology Metrics. http://www.dtic.mil/dtic/kostoff/index.html.

Kostoff, R. N. (1999) "Science and Technology Innovation". Technovation. 19:10. 593-604. October 1999.

Kwok, K. (1995) A Network Approach to Probabilistic Information-Retrieval. ACM Transactions on Information Systems. 13:3.

Lesk, M. (1969) Word-word Associations in Document Retrieval Systems. American Documentation. 20.

Leydesdorff, L. (1987). Various Methods for the Mapping of Science. Scientometrics. 11.

Libbey, M. and Zaltman, G. (1967). The Role and Distribution of Written Informal Communication in Theoretical High Energy Physics. American Institute of Physics, New York.

Lootsma-F.A. (1988). Stochastic and Fuzzy PERT. Technische Hogeschool Delft (Netherlands). Dept. of Mathematics and Informatics Computer Science, Report Number REPT8826, ETN8994007, 25p.

Lotka, A. J. (1926) The Frequency Distribution of Scientific Productivity. Journal of the Washington Academy of Sciences. 16.

MacRoberts, M., and MacRoberts, B. (1996). Problems of Citation Analysis. Scientometrics. 36:3. July-August.

Maron, M. and Kuhns, J. (1960) On Relevance, Probabilistic Indexing, and Information Retrieval. Journal of the ACM. 7.

Melin, G. and Persson, O. (1996). Studying Research Collaboration Using Co-Authorships. Scientometrics. 36:3.

Narin, F. (1989). The Impact of Different Modes of Research Funding. in: Evered, David and Harnett, Sara, Eds. The Evaluation of Scientific Research. John Wiley and Sons, Chichester, UK.

Perko, J. S. and Narin, F. (1997). The Transfer of Public Science to Patented Technology: a Case Study in Agricultural Science. The Journal of Technology Transfer. 22:3.

Peters, H. and Van Raan, A. (1991). Co-Word Based Science Maps of Chemical Engineering. Research Report to the Netherlands Foundation for Technological Research (CWTS-91-03).

Rasmussen, E. (1992) Clustering Algorithms. in Frakes, W. B., and Baeza-Yates, R., (Eds), Information Retrieval: Data Structures and Algorithms. Englewood Cliffs, NJ: Prentice Hall, 1992.

Rip, A. and Courtial, J.P. (1984). Co-word Maps of Biotechnology: An Example of Cognitive Scientometrics. Scientometrics. 6:6.

Robertson, S. and Sparck Jones, K. (1976) Relevance Weighting of Search Terms. JASIS. 27.

Rocchio, J. (1971) Relevance Feedback in Information Retrieval. The Smart System-Experiments in Automatic Document Processing. Prentice Hall, Inc. Englewood Cliffs, NJ.

Salton, G., Fox, E., and Vorhees, E. (1985) Advanced Feedback Methods in Information Retrieval. JASIS. 36.

Salton, G. (1971) Relevance Feedback and the Optimization of Retrieval Effectiveness. The Smart System-Experiments in Automatic Document Processing. Prentice Hall, Inc. Englewood Cliffs, NJ.

Salton, G. and Buckley, C. (1990) Improving Retrieval Performance by Relevance Feedback. JASIS. 41:4.

Seyedghasemipour-S.J. (1987). Stochastic Approach to Project Planning in an R and D Environment: Final Report. East Carolina Univ. Greenville, NC. Dept. of Mathematics, Report Number DOEBC108282, 105p.

Shrum, W. and Mullins, N. (1988). Network Analysis in the Study of Science and Technology. in: Van Raan, A.F.J. ed. Handbook of Quantitative Studies of Science and Technology. North Holland.

Small, H. (1998) A General Framework for Creating Large-scale Maps of Science in Two or Three Dimensions: The SciViz System. SCIENTOMETRICS. 41:(1-2). 125-133. JAN-FEB.

Smeaton, A. and Van Rijsbergen, C. (1983) The Retrieval Effects of Query Expansion on a Feedback Document Retrieval System. Computer Journal. 26.

Spink,A. (1995) Term Relevance Feedback and Mediated Database Searching - Implications for Information-Retrieval Practice and Systems-Design. Information Processing & Management. 31:2.

Stiles, H. (1961) The Association Factor in Information Retrieval. Journal of the ACM. 8.

Swanson, D. R. (1986) Fish Oil, Raynaud's Syndrome, and Undiscovered Public Knowledge. Perspectives in Biology and Medicine. 30:1.

Tijssen, R.J.W. and Vanraan, A.F.J. (1994). Mapping Changes in Science and Technology - Bibliometric Cooccurrence Analysis of the R-and-D Literature. Evaluation Review. 18:1, pp 98-115.

Turner, W.A. (1988). Chartron, G. Laville, F. and Michelet, B. Packaging Information for Peer Review: New Co-Word Analysis Techniques. in: Van Raan, A.F.J. ed. Handbook of Ouantitative Studies of Science and Technology. North Holland.

Van Raan, A.F.J. (1989). Evaluation of Research Groups. in: Evered, David and Harnett, Sara, Eds. The Evaluation of Scientific Research. John Wiley and Sons, UK.

Van Raan, A. (1996). Advanced Bibliometric Methods as Quantitative Core of Peer Review Based Evaluation and Foresight Exercises. Scientometrics. 36:3.

Van Raan, A. and Tijssen, R. (1991). The Neural Net of Neural Network Research: An Exercise in Bibliometric Mapping. Centre for Science and Technology Studies, University of Leiden.

Xu, J. and Croft, W.B. (1996). Query Expansion Using Local and Global Document Analysis. in Proceedings of the 19th International Conference on Research and Development in Information Retrieval (SIGIR'96). Zurich, Switzerland.

Zurcher, R. J. and Kostoff, R. N. (1997). Modeling Technology Roadmaps. Journal of Technology Transfer. 22:3. Fall.